HOCHSCHULE MITTWEIDA
UNIVERSITY OF APPLIED SCIENCES

BACHELORARBEIT

Frau
Bianca Liebscher

Identifikation humanpathogener Pilze auf der Grundlage von qualitätsüberprüften DNA-Sequenzen

Mittweida, 2011

Fakultät
Mathematik, Naturwissenschaft, Informatik

BACHELORARBEIT

Identifikation humanpathogener Pilze auf der Grundlage von qualitätsüberprüften DNA-Sequenzen

Autorin:
Bianca Liebscher

Studiengang:
Biotechnologie/Bioinformatik

Seminargruppe:
BI08w1-B

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Dr. Werner Brabetz

Einreichung:
Mittweida, 24. August 2011

Verteidigung/Bewertung:
Mittweida, 2011

Faculty

Mathematics, Sciences, Informatics

BACHELORTHESIS

Identifcation of humanpathogen fungi based on quality-verified DNA-Sequences

author:

Bianca Liebscher

course of studies:

Biotechnologie/Bioinfomatics

seminar group:

BI08w1-B

first examiner:

Prof. Dr. rer. nat. Dirk Labudde

second examiner:

Dr. Werner Brabetz

submission:

Mittweida, 24. August 2011

defence/ evaluation:

Mittweida, 2011

Bibliografische Beschreibung:

Liebscher, Bianca:
Identifikation humanpathogener Pilze auf der Grundlage von qualitätsüberprüften DNA-Sequenzen. - 2011. - Verzeichnis: 16 Seiten, Inhalt: 75 Seiten, Anhang: 15 Seiten
Mittweida, Hochschule Mittweida, Fakultät Mathematik/Naturwissenschaft/Informatik, Bachelorarbeit, 2011

Referat:

Diese Arbeit beschäftigt sich mit der Formulierung einer Strategie zur Identifizierung humanpathogener Pilze. Sie soll so formuliert werden, dass sie als Grundlage für die Automatisierung und Entwicklung einer entsprechenden Software dienen kann. Weiterhin beschäftigt sich ein Großteil dieser Arbeit mit der Qualität von DNA-Sequenzen, wie sie bestimmt werden und welche Fehler auftreten können.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AC	**Accession Number**
BLAST	**Basic Local Alignment Search Tool**
BOL	**Barcode of Life (Barcode des Lebens)**
BOLD	**Barcode of Life Database**
bp	**Basenpaare**
CBOL	**Consortium for Barcode of Life**
COI, COX1	**mitochondriale Cytochromoxidase Untereinheit I**
DBWG	**Database Working Group**
DDBJ	**DNA Data Bank of Japan**
DNA	**Desoxyribonucleinsäure**
dsDNA	**double stranded DNA (doppelsträngige DNA)**
ssDNA	**single stranded DNA (einzelsträngige DNA)**
dNTP	**DeoxyNukleosidTriPhosphat**
ddNTP	**DiDeoxyNucleosidTriPhosphat**
EPG	**Elektropherogramm**
EMBL	**European Molecular Biology Laboratory**
iBOL	**International Barcode of Life Project**
INSDC	**International Nucleotid Sequence Database Collaboration**
ITS	**Internal Transcribed Spacer**
NCBI	**National Center for Biotechnology Information**
nt	**Nukleotid**
PCR	**polymerase chain reaction (Polymerase-Ketten-Reaktion)**
rDNA	**Ribosomale DNA**

Danksagung

Ich möchte mich bei meinen Betreuern, Prof. Dirk Labudde und Dr. Werner Brabetz für die Bereitstellung des interessanten und vielfältigen Themas sowie für die vielen hilfreichen Ratschlägen und Hinweisen während der Erstellung dieser Arbeit ganz herzlich brdanken.

Des Weiteren danke ich meiner Familie, durch deren Unterstützung das Studium um einiges erleichtert wurde.

Meinem Freund und meinen Kommilitonen möchte ich außerdem meinen Dank aussprechen für die vielen anregenden Gespräche und Denkanstöße.

1 Einleitung

1.1 Motivation

In der Pilzdiagnostik gibt es bisher nur zwei etablierte Methoden zur artspezifischen Identifizierung von Proben: Zum einen kann dies über die morphologische Untersuchung von Pilzkulturen geschehen und zum anderen ist eine Sequenzierung spezifischer Sequenzen sowie ein anschließender Vergleich mit der Datenbank möglich. Doch diese Verfahren sind entweder sehr langwierig (Kultivierung von Pilzen dauert oft Wochen), benötigen umfangreiches Fachwissen, können nicht immer die genaue Art des Pilzes bestimmen oder sind störanfällig bei Verunreinigungen bzw. sind erst gar nicht durchführbar (eine Sequenzierung kann fehlschlagen oder qualitativ schlechte Rohdaten erzeugen wodurch keine Übereinstimmungen in Datenbanken gefunden werden können) [Strandhagen, 2010]. Durch den Einsatz eines Microarrays ist jedoch eine schnelle, einfache und speziesgenaue Identifizierung möglich. Außerdem können Mischproben analysiert werden, ohne vorher eine Isolierung der einzelnen Pilzarten vornehmen zu müssen [Strandhagen, 2010], [Nölte, 2002]. Dies ist vor allem im klinischen Bereich vorteilhaft, da Krankheitserreger schneller diagnostiziert und Patienten schneller mit entsprechenden Medikamenten versorgt werden können. Zum Beispiel könnte eine solche Methode die Diagnose von Mykosen, welche zu den häufigsten Infektionskrankheiten weltweit zählen, stark vereinfachen. Die Anwendung solcher Verfahren ist jedoch nicht auf Krankenhäuser begrenzt, denn als Infektionsquelle werden oft öffentliche Einrichtungen, wie Schulen, Bäder oder Saunen sowie Küchen oder Schlafzimmer in Privaträumen genannt. Die Krankheitserreger können durch molekularbiologisch basierte Methoden besser aufgespürt und unschädlich gemacht werden [Rauch, 2007].

Aus diesem Grund startete die *Biotype Diagnostics GmbH* in Dresden ein Projekt, dessen Ziel es ist, ein Diagnostik-Kit zur Identifikation humanpathogener Pilze zu entwickeln. Als Ausgangspunkt sollen art- und gattungsspezifische DNA-Sequenzen verschiedener Genomregionen dienen.

1.2 Zielstellung

Da ein solches Projekt eine Vielzahl von Arbeitsschritten benötigt, diese manchmal wiederholt, angepasst oder korrigiert werden müssen, ist ein „Arbeitsplan“ von großer Bedeutung. Daher wurde eine Strategie erarbeitet, mit deren Hilfe es möglich sein soll, die Problemstellungen zu lösen, mögliche Fehlerquellen frühzeitig zu erkennen und zu beheben. Die entwickelte Strategie sollte so formuliert werden, dass sie als Grundlage für die Automatisierung und Entwicklung einer entsprechenden Software dienen kann. Weiterhin beschäftigt sich ein Großteil dieser Arbeit mit der Qualität von DNA-Sequenzen, wie sie bestimmt werden und welche Fehler auftreten können. Des Weiteren sollen mit Hilfe eines vorliegenden Datensatzes Kriterien für die Klassifizierung von Sequenzen nach ihrer Qualität entwickelt werden. Außerdem wird anhand eines vorhandenen Datensatzes überprüft, ob die Daten für die Generierung art- und gattungsspezifischer Sequenzen geeignet sind.

1.3 Kapitelübersicht

Für die Analyse von DNA-Sequenzen hinsichtlich ihrer Qualität muss zuerst der Prozess der Sequenzierung erläutert werden. Dies ist notwendig um zu verstehen, wie DNA-Sequenzen bestimmt werden und welche Fehler auftreten können. Im zweiten Kapitel wird das Ausgabe-Format einer Sequenzierung – das Elektropherogramm – sowie deren qualitative Bewertung durch den PHRED-Algorithmus beleuchtet. Des Weiteren gibt dieses Kapitel eine kurze Übersicht über die Qualität von Pilz-Sequenzen in öffentlich zugänglichen Datenbanken. Der dritte Abschnitt beschäftigt sich mit dem DNA-Barcoding als besondere Form der Gewinnung hochqualitativer und standardisierter DNA-Sequenzen. Anschließend erfolgt die Erläuterung der entwickelten Strategie. Das letzte Kapitel beschäftigt sich mit einem wichtigen Teilaspekt der vorgestellten Strategie: der Bewertung von DNA-Sequenzen und deren Rohdaten hinsichtlich ihrer Genauigkeit und Qualität. Eine anschließende Diskussion und ein Ausblick schließen diese Arbeit ab.

2 Sequenzierung

Die Bestimmung von DNA-Sequenzen wird mit Hilfe der Sequenzierung durchgeführt. Es existieren verschiedene Methoden, jedoch wird am häufigsten die Dideoxy- bzw. Kettenabbruch-Methode, welche im Jahr 1977 von Frederick Sanger entwickelt wurde, genutzt [Sensen, 2002 a]. Durch neue technische Möglichkeiten wurde diese Methode hinsichtlich Zeit, Effektivität und Kosten optimiert. Das schließt verschiedene Modifikationen, wie zum Beispiel die Verwendung von fluoreszierenden Farbstoffen für die Detektion statt radioaktiv markierter Stoffe oder die Automatisierung der Reaktion, ein [Sensen, 2002 a], [Nelson, 2005]. Dies hat zum Beispiel zur Entwicklung der *Cycle Sequencing* Technik geführt, welche an die Protokolle einer Polymerase-Ketten-Reaktion[1] angelehnt sind [Heiner, 1998]. Die Länge einer DNA-Sequenz, die mit der Methode nach Sanger bestimmt werden kann, beträgt zwischen 500 und 800 Nukleotiden[2] [Cheng, 2008].

2.1 Dideoxy-Methode nach Sanger

2.1.1 *Verlauf des Prozesses*

Die Kettenabbruchmethode nutzt die Fähigkeiten der DNA-Polymerase mit Hilfe eines Templates[3] DNA-Stränge zu synthetisieren. Für den Start der Synthese ist das Binden eines Primers[4] an das Template erforderlich (vgl. Abbildung 2-2, Nummer 1). Die Zugabe von DNA-Polymerasen und freien Nukleotiden (vgl. Abbildung 2-2, Nummer 2) ermöglichen nun die Erweiterung des DNA-Stranges, ausgehend von der Position der Primer-Hybridisierung. Dabei werden die zum Template komplementären Nukleotide miteinander verknüpft, in dem zwischen der Hydroxylgruppe am 3‘-Ende eines dNTP[5] und der 5‘-Ende des nächsten dNTP eine Phosphodiesterbindung ausgebildet wird. Bei der Kettenabbruchmethode werden dem Reaktionsgemisch jedoch zusätzlich ddNTPs[6]

[1] Abgekürzt als PCR für *Polymerase Chain Reaction*. Dieses Verfahren wird zur Vervielfältigung (Amplifikation) von DNA genutzt.
[2] wird mit „nt“ abgekürzt
[3] einzelsträngige DNA (ssDNA), welche die Abfolge der Nukleotide des zu synthetisierenden DNA-Stranges vorgibt. Wird auch als Matrize bezeichnet. Kann aus einem Zelllysat oder vorangehenden Amplifizierungen mittels PCR gewonnen werden.
[4] ssDNA mit ca. 20-30 nt Länge
[5] DeoxyNukleosidTriPhosphat
[6] DiDeoxyNukleosidTriPhosphat

beigemischt, welche keine Hydroxylgruppe am 3‘-Ende besitzen, wodurch die Ausbildung dieser Verbindung nicht möglich ist und der Strang nicht erweitert werden kann [Sensen, 2002 a]. Der strukturelle Unterschied zwischen dNTP und ddNTP ist in Abbildung 2-1 aufgezeigt. Als Nucleosid wird die glykosidische Verbindung einer Pentose (Zuckermolekül) mit jeweils einer der vier Basen Adenin, Guanin, Cytosin oder Thymin (jeweils mit A, G, C und T abgekürzt) bezeichnet.

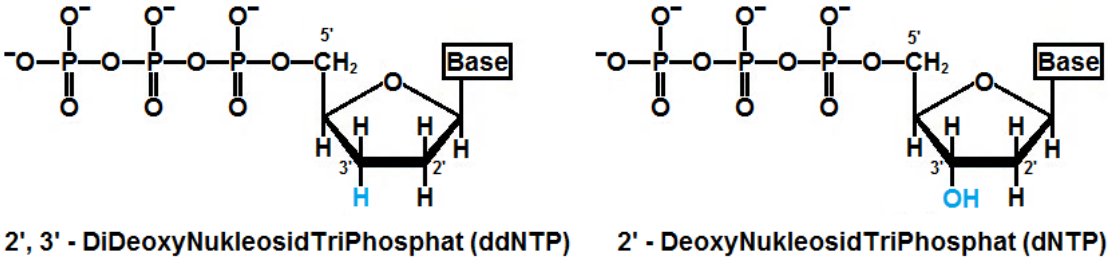

Abbildung 2-1: Vergleich zwischen dNTP und ddNTP

Um zwei Nukleotide miteinander zu verknüpfen wird am 3‘-Ende des Nukleotids mit dem 5‘-Ende des nächsten eine Phosphodiesterbindung ausgebildet. Aufgrund der Abwesenheit einer Hydroxylgruppe bei einem ddNTP ist diese Reaktion nicht möglich und die Erweiterung des DNA-Stranges wird somit abgebrochen [Abbildung nach Nelson, 2005].

Stehen dNTP und ddNTP in einem bestimmten Verhältnis zueinander, entstehen DNA-Fragmente unterschiedlicher Länge, wobei an jedem Nukleotid der zu bestimmenden Sequenz, Kettenabbrüche auftreten [Cheng, 2008]. Ausgehend von der gewählten Markierung (vgl. Kapitel 2.2) können die Reaktionen der vier ddNTP (= ddATP, ddCTP, ddTTP, ddGTP) in einem Reaktionsansatz oder jeweils getrennt erfolgen (in Abbildung 2-2 als getrennte Reaktion dargestellt) [Cheng, 2008]. Nach der Synthesereaktion wird eine Denaturierung[1] mit anschließender Gel- oder Kapillar-Elektrophorese[2] durchgeführt. Dies dient dazu, die DNA - Fragmente ihrer Größe nach aufzutrennen und eine Detektion (vgl. Abbildung 2-2, Nummer 5) der genauen Basenabfolge zu ermöglichen. Dabei muss das Gel so beschaffen sein, dass auch die geringen Größenunterschiede von ein oder zwei Basen deutlich erkennbar sind (vgl. Abbildung 2-2, Nummer 4) Durch eine Detektion der Banden werden Elektropherogramme erstellt [Nelson, 2005].

[1] Die zwei Stränge der dsDNA werden mit Hilfe von Temperaturerhöhungen voneinander getrennt und in ssDNA überführt.

[2] Die DNA wird auf ein Agarose- oder Acrylamidgel aufgetragen. Durch elektrische Spannung wandern kleine DNA-Fragmente schneller als große durch das Gel und es bilden sich Banden. Eine Bande beinhaltet Fragmente gleicher Länge.

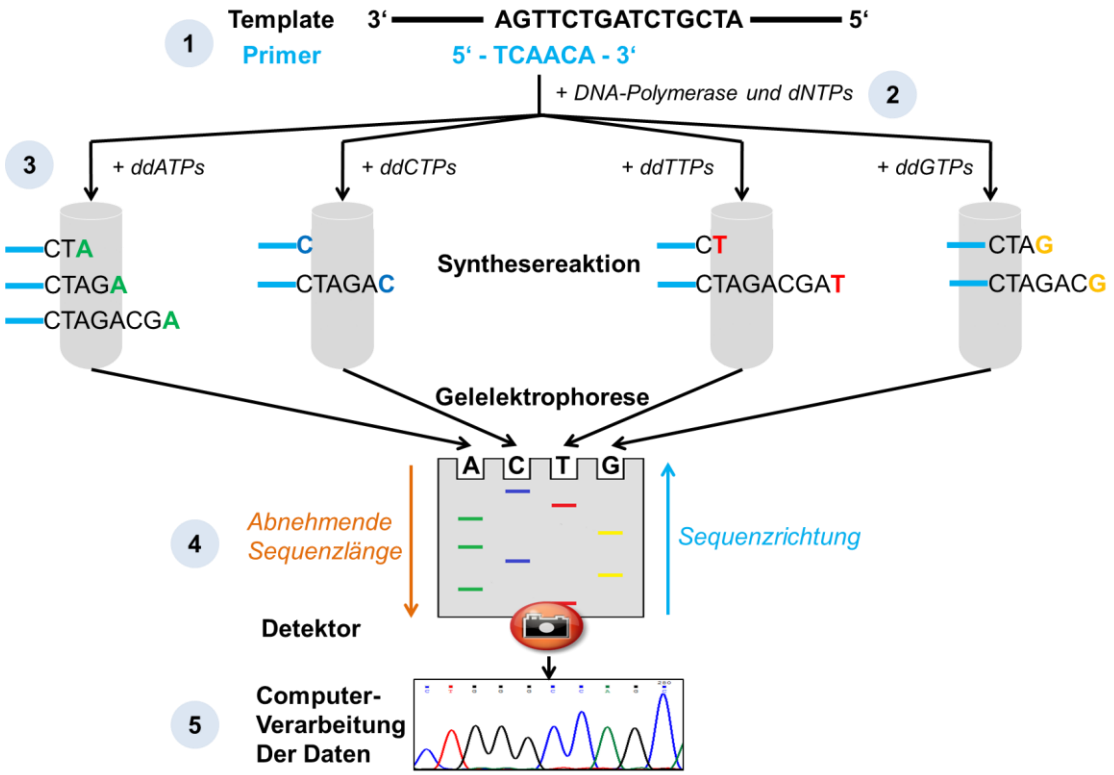

Abbildung 2-2: DNA Sequenzierung nach Sanger

In einem Reaktionsgemisch liegen Template und Primer vor (1). Durch Zugabe von DNA-Polymerasen und dNTPs (2) sowie jeweils eines der vier fluoreszenzmarkierten ddNTPs zu je einem Ansatz startet die Synthesereaktion (3). Es entstehen DNA-Fragmente unterschiedlicher Länge, welche anschließend durch eine Gelelektrophorese aufgetrennt (4) und die Fluoreszenzsignale mit einem Detektor aufgenommen werden. Es entstehen so genannte Elektropherogramme, welche durch computeranalytische Software in die DNA-Sequenz konvertiert werden (5).

2.1.2 *Parallele bidirektionale Sequenzierung*

Diese Methode beschreibt die gleichzeitige Bestimmung von zwei komplementären DNA-Strängen in einer Sequenzierung. Ähnlich einer PCR, ist es notwendig zwei verschiedene Primer einzusetzen. Diese werden als Forward- oder Reverse-Primer bezeichnet und sind mit jeweils einem bestimmten fluoreszierenden Farbstoff markiert. Ein Primer bindet auf dem Matrizen-Strang und der zweite auf dem Folgestrang eines dsDNA-Templates. Die Synthese der einzelnen DNA-Fragmente wird in vier verschiedenen Ansätzen durchgeführt (vgl. Kap.:2.2.2). Jedoch werden pro Ansatz nun Fragmente von zwei DNA-Strängen parallel hergestellt und zur Detektion in einer Gel-Spur elektrophoretisch aufgetrennt. Um im Gel die Fragmente dem jeweiligen Template-Strang zuordnen zu können, werden die unterschiedliche markierten Primer von zwei verschiedenen Laser zum Fluoreszieren angeregt. Zwei Detektoren, welche

am Ende des Gels positioniert sind, zeichnen die zwei unterschiedlichen Fluoreszenzsignale auf (vgl. Abbildung 2-3). Als Ausgabe erhält man zwei komplementäre DNA-Sequenzen [Wiemann, 1995], [Sensen, 2002 b].

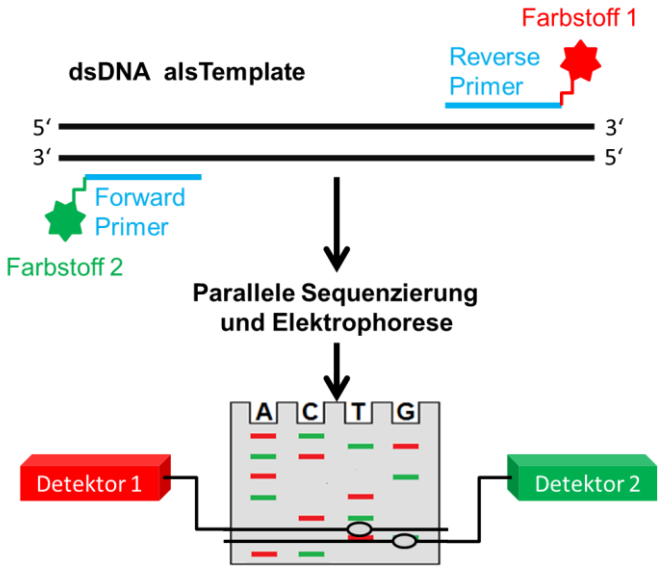

Abbildung 2-3: bidirektionale Sequenzierung

Mit Hilfe dieser Technik können beide Stränge einer dsDNA in einer Sequenzierungs-Reaktion bestimmt werden. Vorteile dieser Methode ist die Bestimmung einer Sequenz mit doppelter Genauigkeit, sowie geringerer Materialverbrauch und Zeitersparnis.

Die Vorteile dieser technischen Erweiterung sind vielfältig. Zum einen kann eine DNA-Sequenz mit doppelter Genauigkeit bestimmt und eventuelle Fehler leichter behoben werden. Zum anderen sinkt der Material-, Zeit- und Geld-Verbrauch, da zwei Reaktionen parallel erfolgen. Des Weiteren wird mit dieser Methode die Anzahl der Sequenzinformation, d. h. die bestimmbare Menge an Nukleotiden, verdoppelt.

2.2 Markierung der DNA-Stränge

In der Originalarbeit von Sanger wird die Markierung der DNA noch mit Hilfe von radioaktiven Isotopen durchgeführt. Aufgrund der gesundheitlichen Gefahren und fehlenden Möglichkeiten zur Automatisierung des Prozesses werden heute fluoreszierende Stoffe als Marker benutzt [Sensen, 2002 a]. Dabei unterscheidet man zwischen zwei Verfahren: der „*dye terminator chemistry*“ und der „*dye primer chemistry*“ [Ewing, 1998 a].

2.2.1 *„Dye terminator chemistry“*

Bei dieser Methode werden vier verschiedene Fluorophore, die in unterschiedlichen Wellenlängen Licht emittieren, verwendet. Jedes der vier ddNTPs wird nun mit einem Farbstoff verbunden (z.B. ddATP grünes, ddCTP blaues, ddGTP gelbes und ddTTP rotes Fluorophor). Dies erlaubt die Durchführung der Reaktion in einem einzigen Ansatz und Auftrennung der Fragmente in einer einzigen Spur des Gels. Passiert eine DNA-Bande den Detektor, welcher am Ende des Gels positioniert ist, wird ein basenspezifisches Signal aufgezeichnet. Aus der Abfolge der jeweiligen Signalwerte kann die DNA-Sequenz geschlussfolgert werden (vgl. Abbildung 2-2, Nummer 5) [Cheng, 2008].

2.2.2 *„Dye primer chemistry“*

Alternativ zu der Markierung der ddNTPs können auch Primer mit Farbstoffen verbunden und zur Detektion genutzt werden. Bei dieser Variante muss die Sequenzierung jedoch in vier getrennten Reaktionsansätzen bzw. Gelspuren erfolgen, da keine basenspezifische Zuordnung der Signale möglich ist. Jeder Ansatz enthält daher nur eines der vier ddNTP (z.B. nur ddCTPs) [Cheng, 2008].

2.3 Detektion

Die Erschließung der Sequenzdaten ist bei der manuellen Sequenzierung mit radioaktiven Isotopen erst dann möglich, wenn die DNA-Banden eines Elektrophorese-Gels mittels Autoradiographie sichtbar gemacht wurden [Cheng, 2008]. Dieses aufwendige Verfahren konnte durch Verwendung fluoreszierender Farbstoffe bei der automatischen Sequenzierung ersetzt werden. Es ist damit möglich die DNA-Banden in Echtzeit, d.h. während der elektrophoretischen Auftrennung, zu detektieren [Primrose, 2006]. Ein Laser, welcher am unteren Rand des Gels positioniert ist, regt die in den Banden enthaltenen Fluorophore an. Daraufhin beginnen diese Licht in einer bestimmten Wellenlänge auszusenden. Der Detektor zeichnet nun die Intensitäten der Lichtemissionen auf und gibt sie an einen Rechner weiter [Ewing, 1998 a]. Diese, als Rohdaten (*„raw data“*) bezeichneten, Aufnahmen bestehen aus einem vier-kanaligem digitalen Signal. Jeder Kanal ist einem bestimmten fluoreszierendem Farbstoff und dem entsprechendem Nukleotid zugeordnet. Werden diese vier Kanäle durch spezielle Software visualisiert entsteht ein Diagramm, welches als Chromatogramm, Elektropherogramm[1] oder Trace (vgl. Abbildung 2-4) bezeichnet wird [Cheng, 2008], [Wendl, 2001].

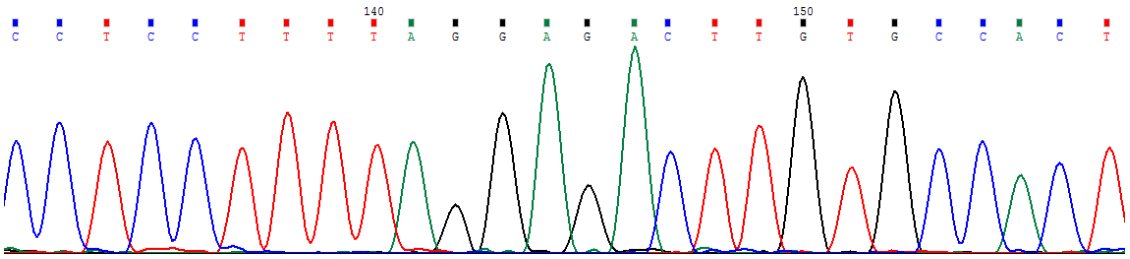

Abbildung 2-4: Elektropherogramm einer automatischen Sequenzierung

Als Elektropherogramm werden die durch eine automatische Sequenzierung gewonnenen Daten bezeichnet. Die Kurven geben den Verlauf der Intensitäten der Fluoreszenz-Signale bei der Elektrophorese wieder. Je höher ein Peak desto stärker das Signal. Aus der Abfolge der farbcodierten Signale kann man die DNA-Sequenz schlussfolgern.

Der Prozess der Konvertierung der Signalabfolge in eine DNA-Sequenz wird als *Base-Calling* bezeichnet. Dabei spielen verschiedene Parameter wie Länge, relative Höhe oder Lage eines Peaks sowie das Signal/Rausch- Verhältnis eine große Rolle (siehe Kapitel 3) [Merkl, 2009 b].

[1] wird mit EPG abgekürzt

3 Qualität von Sequenzdaten

Die Bestimmung von DNA-Sequenzen mittels Sequenzierung verläuft nicht immer fehlerfrei und die Interpretation von EPGs ist abhängig von Rohdaten und Algorithmus. Daher ist die Bewertung einer Sequenz hinsichtlich ihrer Qualität und Genauigkeit notwendig um abschätzen zu können, ob eine Sequenz für weitere Analysen genutzt werden kann oder nicht. Ein weiterer wichtiger Aspekt ist die richtige Annotation einer Sequenz bei der Veröffentlichung in einer Datenbank, um dem Nutzer möglichst genaue Informationen bereitzustellen.

3.1 Grundlagen der Elektropherogramme

3.1.1 *Elektronische Verarbeitung der Rohdaten*

Die bei der Detektion aufgenommenen Fluoreszenzsignale können als Abbild des Gels betrachtet werden. Die Konvertierung dieses Abbilds in eine DNA-Sequenz wird in vier Schritten durchgeführt. Abhängig von der verwendeten Elektrophorese-Methode werden im ersten Schritt, dem *Lane Tracking,* die Banden des Gels identifiziert[1]. Danach wird jedes der vier Signale aufsummiert um ein Profil (Trace) zu erstellen. Dieses besteht aus vier Arrays, in denen die Signalintensitäten in Abhängigkeit von der Zeit abgespeichert werden. Dies wird als *Lane Profiling* bezeichnet. Das *Trace Processing* ist für die Glättung und Reduzierung des Rauschens der Signale durch bestimmte Algorithmen zuständig. Der letzte Schritt ist für die Generierung einer genauen und möglichst fehlerfreien DNA-Sequenz aus dem Trace verantwortlich und wird *Base-Calling* genannt [Lawrence, 1993], [Ewing, 1998 a].

3.1.2 *Das ideale Elektropherogramm*

Theoretisch werden die DNA-Fragmente bei der Gelelektrophorese so aufgetrennt, dass eindeutige, klar getrennte Banden entstehen und in Folge dessen scharfe Fluoreszenzsignale detektiert werden. Daraus ergibt sich eine Trace, dessen Peaks sauber getrennt sind, nicht überlappen und jeweils gleiche Abstände zueinander besitzen. Somit kann jeder Peak einem Fragment bestimmter Länge zugeordnet und die DNA-

[1] bei der Kapillarelektrophorese ist dieser Schritt nicht notwendig

Sequenz ohne Probleme aus den Rohdaten abgelesen werden (vgl. Abbildung 3-1) [Wendl, 2001], [Ewing, 1998 a].
Durch Template bedingte Unterschiede beim Einbau von ddNTPs in die Sequenz können die Höhen der Peaks im Elektropherogramm variieren. Zum Beispiel sind „G“-Peaks, die auf einen „A“-Peak folgen, meist niedriger. Diese Effekte konnten jedoch durch die Entwicklung neuer Enzyme und Farbstoffe für die Sequenzierung reduziert werden.

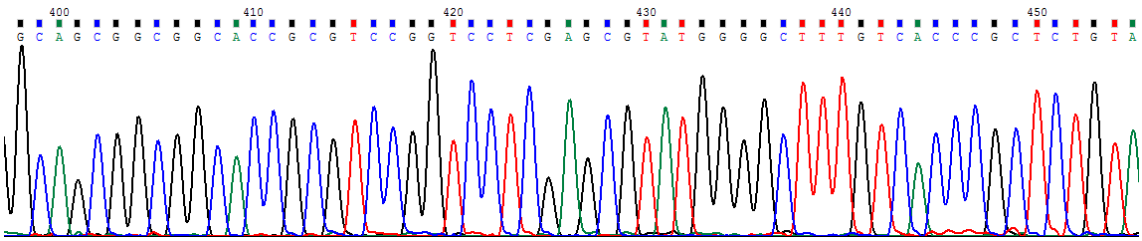

Abbildung 3-1: ideales Elektropherogramm

Der ideale Trace zeigt klar voneinander getrennte, nicht überlappende Peaks aus denen die DNA-Sequenz einfach geschlussfolgert werden kann.

Aufgrund verschiedener Parameter und äußerer Einflüsse, die in jeder Sequenzierung unterschiedlich sein können, variiert auch die Qualität der Rohdaten. Um trotzdem eine möglichst der Realität entsprechende DNA-Sequenz aus den Daten ableiten zu können, muss eine Fehlerbetrachtung durchgeführt werden.

3.1.3 *Das reale Elektropherogramm*

Die qualitativen Unterschiede bestehen nicht nur zwischen DNA-Sequenzen, die mit unterschiedlichen Sequenzierungen bestimmt wurden, sondern können auch innerhalb einer Trace beobachtet werden. Diese sind nicht immer Ursache einer fehlerhaften Sequenzierung, sondern kommen in fast allen Elektropherogrammen vor. Oft sind die ersten 50 Peaks stark verrauscht und werden beim *Base-Calling* falschen Nukleotiden zugeordnet. Dies ist auf die irreguläre Migration durch sehr kurze Fragmente im Gel zu begründen. Des Weiteren können freie farbstoffhaltige Moleküle vom Detektor erfasst werden und so zu falschen Signalen führen (vgl. Abbildung 3-2) [Ewing, 1998 a], [Wendl, 2001].

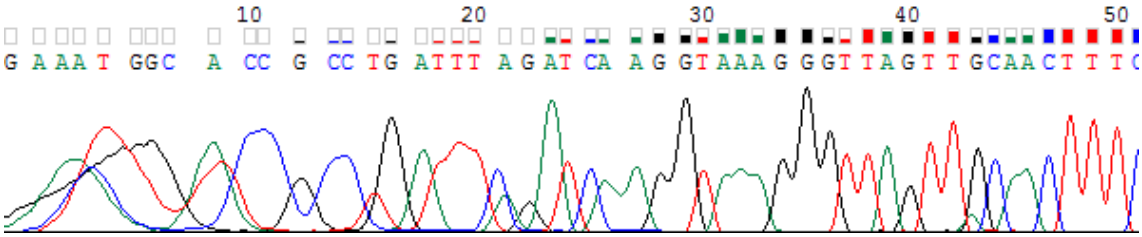

Abbildung 3-2: schlechte Qualität am Anfang eines EPG

Überlagerte, unterschiedlich breite sowie hohe Sekundär-Peaks verringern die Sequenzqualität und erschweren das Ablesen der Sequenz. Grund dafür sind unterschiedliche Wanderungsgeschwindigkeiten von kleinen Fragmenten zu Beginn der Elektrophorese.

Die Peaks in der Mitte, d.h. der Bereich von ca. 50-600 nt, sind meist von sehr guter Qualität und sind abschnittsweiseweise mit einem idealen Trace vergleichbar. Ist dies nicht der Fall, so deutet dies auf Prozess-Fehler hin (vgl. Abbildung 3-1).

Am Ende des EPGs sinkt die Qualität der Peaks wieder. Grund dafür sind steigende Diffusionseffekte, abnehmende relative Massendifferenzen zwischen den größer werdenden DNA-Fragmenten und die sinkende Anzahl markierter Moleküle. Dadurch werden die Fluoreszenzsignale schwächer und die Peaks sind nicht mehr eindeutig zuzuordnen. In diesem Abschnitt des Trace kommt es ebenfalls häufig zu einem plötzlichen Signalverlust. Dies tritt dann auf, wenn eine PCR vor der Sequenzierung durchgeführt wurde und die Proben-DNA kürzer ist, als die durch Sequenzierung maximal bestimmbare Sequenzlänge (vgl. Abbildung 3-3) [Ewing, 1998 a], [Wendl, 2001].

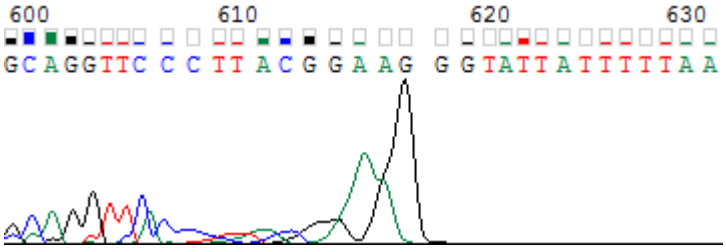

Abbildung 3-3: schlechte Qualität am Ende eines EPG

Im hinteren Bereich eines EPG (ca. ab der 600.-700. Base) sind die Peaks wieder schlechter getrennt und die Signalstärken nehmen stetig ab. Plötzliche Signalverluste treten vor allem dann auf, wenn vor der Sequenzierung eine PCR durchgeführt wurde.

3.2 Ursachen für Fehler in Traces

Die Gründe für qualitativ schlechte Elektropherogramme sind vielfältig. Die folgende Aufzählung gliedert die möglichen Fehlerquellen nach den verschiedenen Teilprozessen der Sequenzierung und gibt einen Überblick über die häufigsten Ursachen für fehlgeschlagene Sequenzierungen. Da sowohl in Kapitel 3.2.1 als auch 3.2.2 biologische Komponenten für die jeweiligen Reaktionen erforderlich sind, werden diese der Übersicht halber in Kapitel 3.2.3 gesondert behandelt.

3.2.1 *Fehler bei DNA-Aufbereitung (z.B. PCR)*

Bevor eine Sequenzierung gestartet wird, muss gesichert sein, dass ausreichend DNA für die Reaktion vorhanden ist. Um dies zu erreichen, kann die DNA durch eine PCR oder bei Genomprojekten mit Hilfe von Vektoren (Klonierung) amplifiziert werden [Achaz]. Durch vergessene oder *unzureichende Aufreinigung* der Reaktionsgemische nach der Reaktion können die PCR-Primer oder Inhibitoren mit in die Sequenzierung eingebracht werden. Des Weiteren ist eine *Kontamination mit Fremd-DNA* durch fehlende Schutzmaßnahmen, wie zum Beispiel nicht sterile Bedingungen im Labor, möglich [Cheng, 2008]. Meist führen Fehler in diesem Schritt zu einer fehlgeschlagenen Sequenzierung oder zu einer sehr schlechten Qualität der Rohdaten.

3.2.2 *Fehler während der Sequenzier-Reaktion*

Die Synthese der DNA-Fragmente kann nur erfolgreich verlaufen, wenn alle benötigten Stoffe in idealer Konzentration vorliegen und die physikalischen Bedingungen, z.B. Temperatur optimal eingestellt sind. Mögliche Ursachen, die die Synthese der DNA-Fragmente verhindern können, sind *Kontaminationen* mit Salzen (wie EDTA), Proteinen (DNasen[1]), organischen Chemikalien (z.B. Ethanol), *Fremd-DNA* oder hohen Konzentrationen an divalenten Kationen (z.B. Kalzium) [Kieleczawa, 2004]. Des Weiteren kann die DNA bei der Aufreinigung *verloren gehen* oder es wird *zu viel DNA* für die Sequenzierung verwendet. Ein weiteres Problem ist die *Lagerung der Reagenzien unter falschen Bedingungen* sowie zu häufiges Auftauen und Einfrieren Dieser. Die Folge sind degradierte DNA-Polymerasen oder Nukleotide [Cheng, 2008]. Einen wichtigen

[1] Sind DNA zersetzende Moleküle.

Faktor stellt auch die Wahl der *korrekten Primer* sowie deren *Annealingtemperatur*[1] dar. Es muss gesichert sein, dass nur eine Primer-Bindungsstelle auf der Template-DNA existiert, um Kontaminationen mit unerwünschter DNA zu vermeiden [Cheng, 2008].

3.2.3 *Fehler durch biologische Komponenten*

Der Begriff biologische Komponenten steht zusammenfassend für alle Primer, DNA-Templates, synthetisierte DNA-Stränge sowie die dafür benötigten DNA-Polymerasen.
Ein wichtiger Faktor, vor allem bei großen Sequenzierprojekten, ist die natürlich vorhandene *Fehlerrate der DNA-Polymerase*. Sie schwankt zwischen 10^{-4} und 10^{-6} Fehler/Basenpaar) und kann zu *Single Nucleotide Polymorphisms* (SNP) führen. Dies bedeutet, dass ein anderer, nichtkomplementärer Nukleotid eingebaut und so eine Substitution herbeigeführt wird [Achaz, 2008], [Chevreux, 2005].
Des Weiteren ist es möglich, dass die Template - DNA von *heterozygoten Regionen* eines diploiden Organismus stammt und dadurch ebenfalls geringe Abweichungen zwischen den Sequenzen entstehen [Nucleics, 2010].
Bei der Amplifizierung der DNA mit Hilfe von Vektoren, kann es vorkommen, dass die Klone durch eine *spontane Rekombination* neu angeordnet werden, indem zwei, normalerweise räumlich voneinander getrennte, Regionen des Genoms aneinandergeknüpft werden. Die neue Sequenz wird als *Chimäre* bezeichnet [Chevreux, 2005].
Ein Abbruch der DNA-Synthese kann herbeigeführt werden, wenn das Template *starke Sekundärstrukturen* (vgl. Abbildung 3-4*)* ausbildet. Dies geschieht vor allem, wenn in der Sequenz ein hoher GC-Gehalt vorliegt [Ewing, 1998 a].

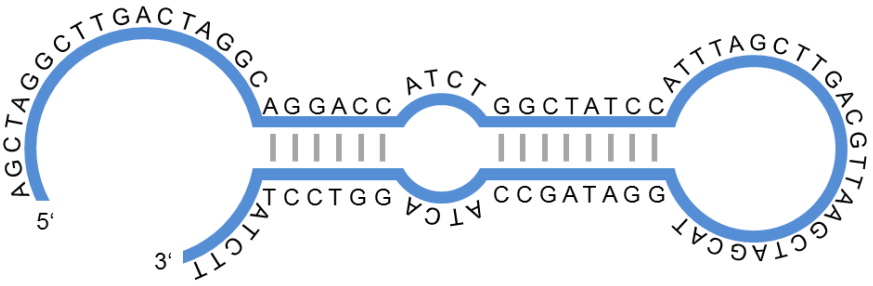

Abbildung 3-4: Sekundärstruktur einer DNA

Sekundärstrukturen, wieder abgebildete *Hairpin-Loop* und *internal bulge* können die DNA-Polymerase daran hindern einen DNA-Strang zu erweitern. Dies führt zum Abbruch der Synthese und zeigt sich im EPG durch plötzlichen Signalverlust.

[1] Die Temperatur bei der die Primer optimal an die DNA binden.

Gegenteilig dazu können lange *Homopolymer-Sequenzen* aus Thymin oder Adenin sowie *short tandem repeats* (STR) mit einer Länge von 2 – 4 bp zu einem Verrutschen der DNA-Polymerase (*Slippage*) auf der DNA um ein paar Nukleotide führen. Dies ist durch die schwächeren Bindungskräfte zwischen Adenin und Thymin (nur zwei Wasserstoffbrückenbindungen) zu begründen. Die Folge sind unterschiedlich lange Sequenzen und damit verschobene Peaks im EPG. Ein Slippage erzeugt im EPG eine wellenartige Anordnung der Peaks [Nucleics, 2010] [Cheng, 2008].

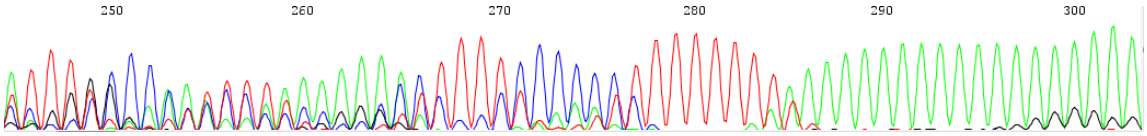

Abbildung 3-5: Wellenartiges Elektropherogramm durch Slippage

Die DNA-Polymerase kann bei langen Mononukleotid-Sequenzen (z.B. achtmal Thymin hintereinander) den halt zum Template verlieren und an einer anderen Stellen neu Binden. das führt zu unterschiedlich langen Sequenzen und einem wellenartigen Trace

Nicht nur das Template kann Sekundärstrukturen bilden, sondern auch *Primer können miteinander bzw. mit sich selbst hybridisieren*, wenn sie falsch ausgewählt worden sind. Dadurch stehen für die Sequenzierung weniger bis keine freie Primer zur Verfügung und es entstehen nur wenige DNA-Fragmente, die detektiert werden können [Cheng, 2008].

3.2.4 *Fehler während der Elektrophorese*

Bei der Auftrennung der DNA-Fragmente ist es wichtig, dass die Wanderungsgeschwindigkeit konstant bleibt und die Banden klar voneinander zu unterscheiden sind. Aufgrund *terminaler Sekundärstrukturen* (z.B. HairPin Loops) kann diese jedoch verändert werden und die Fragmente wandern schneller durch das Gel. Es kommt zu einer Verschiebung bzw. Überlagerung der Banden, die als *Kompressionen* bezeichnet werden. Durch die Benutzung der „*dye terminator chemistry*“, der Denaturierung oder Temperaturerhöhung können diese Effekte jedoch weitestgehend verhindert werden [Lipshutz, 1994], [Ewing, 1998 a], [Wendl, 2001].
Bei kleinen Fragmenten sind die *Effekte zwischen Farbstoff und Sequenz* noch relativ groß, wodurch ebenfalls das Wanderungsverhalten beeinflusst wird. Ungebundene Primer, freie ddNTPs sowie Temperaturschwankungen im Gel können auch Ursache *verschmierter Banden* sein [Ewing, 1998 a], [Wendl, 2001].

Eine schlechte Qualität von EPGs kann auch durch eine *überladene oder blockierte Kapillare* herbeigeführt werden [Cheng, 2008]. Weiterhin können so genannte *Dye-Blobs* in einem Elektropherogramm auftreten. Das sind mehrere Peaks überlagernde, sehr breite und hohe Peaks. Sie können entstehen, wenn nach einer Sequenzierung die fluoreszierenden Farbstoffmoleküle nicht komplett entfernt wurden oder es setzt bei der Speicherung der Sequenzier-Produkte eine langsame Hydrolyse der Farbstoffe ein. Im ersten Fall entsteht im EGP ungefähr an Position 50 und im zweiten Fall um die Position 100 ein T-Blob (Thymin-Blob). [Cheng, 2008] Dye-Blobs können auch entstehen, wenn bei der Aufreinigung der DNA nach der Sequenzierung zu hohe Ethanol Konzentrationen verwendet werden. Die Höhen der Dye-Blobs können ebenfalls stark variieren [Nucleics, 2010].

3.2.5 *Fehler bei der computergestützten Verarbeitung der Daten*

Für die Konvertierung der Elektropherogramme in DNA-Sequenzen wird spezielle Software benötigt. Die Genauigkeit des verwendeten Algorithmus ist dabei ausschlaggebend für die Richtigkeit der DNA-Sequenz. Durch die von Experiment zu Experiment schwankenden Signalstärken, Überlagerungen von Peaks oder starkes Rauschen treten oft Fehlinterpretationen auf. Dies äußert sich in der DNA-Sequenz als Insertionen, Deletionen oder Substitutionen [Chevreux, 2005] [Cheng, 2008]. Des Weiteren müssen die Primer – oder Vektorsequenzen, die am Anfang bzw. am Ende der Sequenz vorhanden sind, entfernt werden. Dies ist nicht durch einfache Sequenzvergleiche möglich, da sie in den qualitativ schlechtesten Bereichen eines EPG lokalisiert sind (vgl. Kap. 3.1.3) und die Nukleotide oft falsch bestimmt werden [Chevreux, 2005].

3.2.6 *Auswirkung von Fehler auf die Sequenz*

Liegt eine aus den EPG geschlussfolgerte Sequenz vor, so kann diese in drei verschiedenen Punkten von der Originalsequenz, wie sie im Organismus vorliegt, abweichen (vgl. Abbildung 3-6).
Erstens: zu breite oder mehrere eng aneinander liegende Peaks können als zwei Nukleotide bewertet werden, was eine Insertion eines Nucleotids zur Folge hat.
Zweitens: Bei der Überlagerung zweier Peaks, durch z.B. Kompressionen, ist das Programm nicht immer in der Lage diese als zwei Nukleotide zu erkennen und es entsteht eine Deletion.

Drittens: Das Vorhandensein eines sekundären Peaks führt oftmals zu einer falschen Zuordnung zwischen Signal und zugehörigem Nukleotid, vor allem wenn dieser mehr als 20 % der Höhe des primären Peaks erreicht. Dies wird als Substitution bezeichnet [Chevreux, 2005].

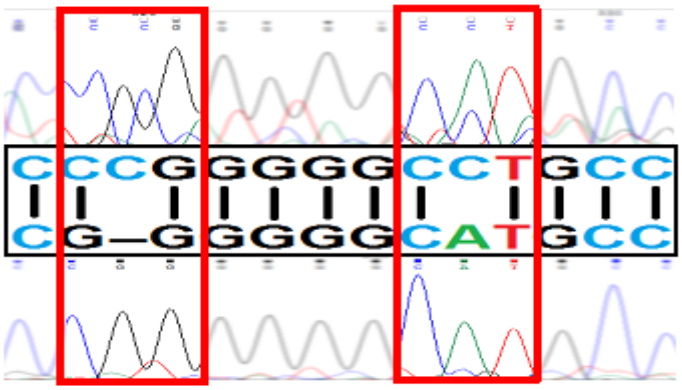

Abbildung 3-6: Auswirkung von Sequenzier-Fehler auf die Sequenz

Zu viele Peaks können als mehrere Nukleotide interpretiert werden. Wird diese Sequenz mit der Originalsequenz verglichen entsteht eine Insertion (linker roter Rahmen). Das Vorhandensein relativ hoher Sekundärer Peaks kann dazu führen, dass ein falscher Nukleotid erkannt wird. Es entsteht eine Substitution (rechter roter Rahmen)

3.3 Qualitätsbewertung von Traces

Um Aussagen darüber treffen zu können, ob eine Sequenz präzise genug ist, wird beim *Base-Calling* nicht nur die Sequenz bestimmt sondern gleichzeitig deren Qualität berechnet. Dafür gibt es verschiedene Ansätze und Algorithmen, die auf positionsspezifischen Fehlerwahrscheinlichkeiten aller Nukleotide einer Sequenz basieren. [Richterich, 1998] Das Programm Phred, welches in Kapitel 3.3.2 näher erläutert wird, liefert die besten Ergebnisse und wird nun als Standardanwendung für Qualitätsbestimmungen verwendet. [Wendl, 2001]

3.3.1 *Praktischer Nutzen*

Die computergestützte Analyse von EPGs erlaubt eine sofortige Qualitätskontrolle von Sequenzen und gibt Aufschluss über Häufigkeit und Lage von Fehlern in Rohdaten. Die noch notwenigen manuellen Korrekturen können so spezifisch, mit einer hohen Zeitersparnis, durchgeführt und der Aufwand eventuell notweniger Neusequenzierungen besser abgeschätzt werden. Des Weiteren ist ein einfaches und schnelles Vergleichen zwischen Sequenzen aus unterschiedlichen Sequenzier - Methoden oder –Projekten möglich [Richterich, 1998].

3.3.2 *Qualitätswerte des Programms Phred*

Der Algorithmus des Programms besteht aus vier Phasen. Mit ihm ist es möglich Fehlerraten verschiedener Sequenzier - Methoden, Maschinen und Parameter zu bestimmen [Wendl, 2001]. In der ersten Phase werden idealisierte Peak-Standorte (vorhergesagte Peaks) mit Hilfe einfacher Fourier - Methoden festgelegt. Dabei wird die Tatsache genutzt, dass die meisten Peaks lokal klar voneinander getrennt sind und die richtige Anzahl von Nukleotiden erkannt wird. Dies sollte möglichst auch in Regionen mit starken Rauschen oder Kompressionen möglich sein. Die Zweite Phase identifiziert die wirkliche Position der Peaks im Trace (beobachtete Peaks).
Im nächsten Schritt werden die vorhergesagten mit den beobachteten Peaks verglichen und es wird versucht, diese aneinander anzupassen. Aus diesem Grund werden manche Peaks nicht betrachtet oder geteilt. Die Abfolge der übereinstimmenden beobachteten Peaks gibt dabei die Abfolge der DNA-Seque nz wieder. Es ist möglich, dass im ersten Schritt Peaks aufgrund von Kompressionen oder starkem Rauschen eine falsche Anzahl vorhergesagter Peaks entstehen. Daher werden manche

beobachteten Peaks nicht verarbeitet und müssen in einem letzten Schritt extra überprüft werden. Dabei muss dieser Peak verschiedene Kriterien erfüllen um als richtig zu gelten. Erst dann kann ein zusätzlicher Nukleotid in die schon bestimmte Sequenz eingefügt werden. Kann zu einem vorhergesagten Peak kein beobachteter Paek zugeordnet werden, wird in der Sequenz ein „N“ eingefügt [Ewing, 1998 a].

Für die Bewertung der Qualität jedes Nukleotids wird ein „Quality Score“ q mit Hilfe der Formel $q = -10 \cdot log_{10}(p)$ berechnet, wobei p die Fehlerwahrscheinlichkeit des Nukleotids darstellt [Ewing, 1998 a]. In Tabelle 3-1 sind einige Werte für q mit entsprechender Genauigkeit des Nukleotids dargestellt. Im Allgemeinen wird Quality Score von 20 (auch Phred20) als Schwellwert für ein „richtiges“ Nukleotid genutzt [Wendl, 2001].

Tabelle 3-1: „Quality Score“ nach Phred [1]

Die Fehlerraten der Nucleotide werden mit Hilfe einer logarithmischen Transformation in *Quality Scores* umgerechnet.

Quality Score	Anzahl falscher Nukleotide	Fehlerrate
10	1 auf 10	10 %
20	1 auf 100	1 %
30	1 auf 1.000	0,1 %
40	1 auf 10.000	0,01 %
50	1 auf 100.000	0,001 %

3.3.3 *Validierung des Quality Scores*

Die Genauigkeit des PHRED-Algorithmus wurde anhand verschiedener statistischer Analysen bewiesen. Es wurde gezeigt, dass die Fehlerraten, die mit PHRED bestimmt wurden, um 39 – 52 % geringer[2] sind, als die Fehlerraten der ABI Analyse Software[3]. Dadurch können längerer Sequenzabschnitte höherer Qualität gewonnen und Neu-Sequenzierungen verhindert werden [Ewing, 1998 b]. Des Weiteren wurden die von PHRED bestimmten Fehlerraten mit den tatsächlichen Werten von sechs Projekten, mit jeweils unterschiedlicher Sequenzier-Methode, verglichen [Richterich, 1998]. Die graphische Auswertung der Ergebnisse ist in Abbildung 3-7 dargestellt. Es sind jeweils

[1] Tabelle entnommen von http://www.phrap.com/phred/
[2] ist abhängig von der verwendeten Markierung und Länge der Sequenzen
[3] ABI Analyse Software ist die Base-Calling Software die für die Sequenzier-Automaten bereitgestellt wird.

die Fehlerraten gegen die Position der Nukleotide aufgetragen. Es ist erkennbar, dass trotz der unterschiedlichen Kurvenverläufe die berechneten Fehlerraten sehr gut mit den wirklichen übereinstimmen. Weiterhin ist anhand der Kurvenverläufe die Qualitätsverteilung in Abhängigkeit von der Position erkennbar. Die Ursachen für die Qualitätsunterschiede innerhalb einer Sequenz wurden in Kapitel 3.1.3 bereits näher beschrieben.

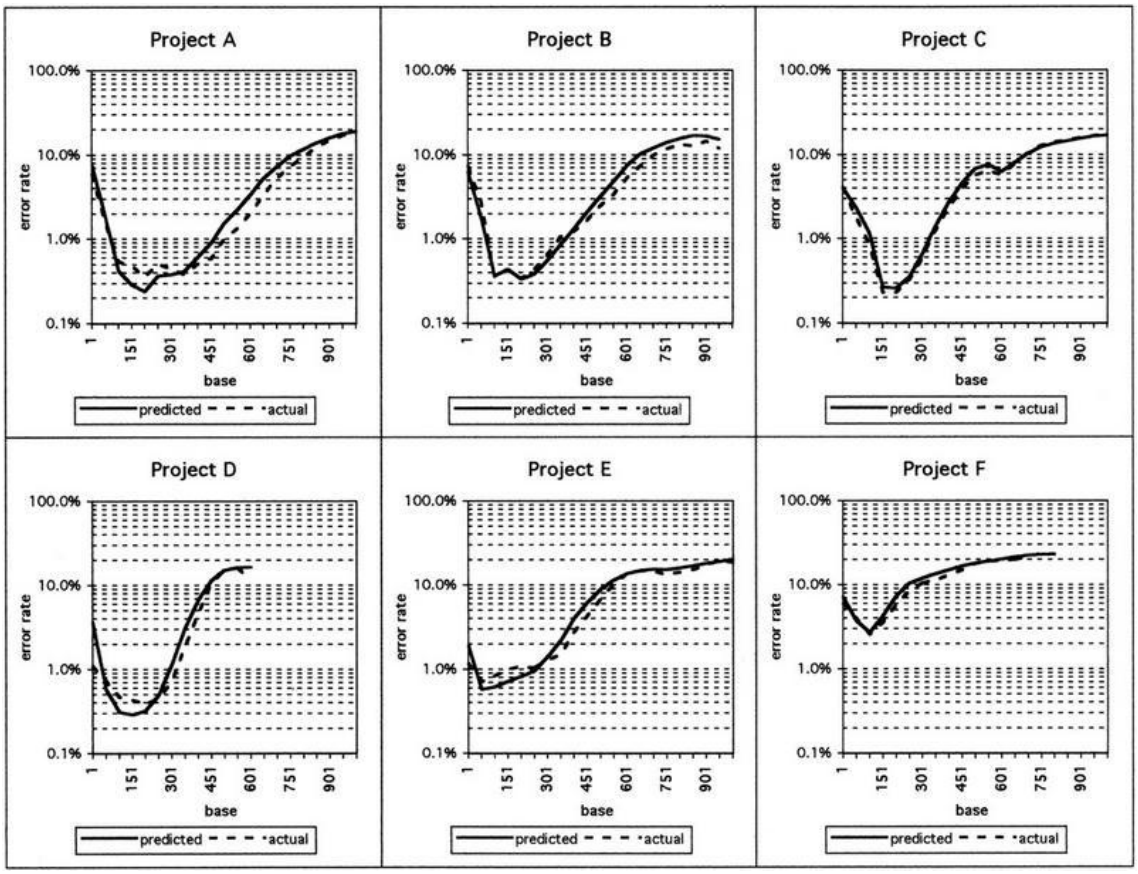

Abbildung 3-7: Vergleich der realen und berechneten Fehlerraten

Ein Vergleich der von PHRED berechneten Fehlerraten (durchgehende Linie) mit den tatsächlichen Fehlerraten (unterbrochene Linie), liefert den Beweis der hohen Genauigkeit des PHRED-Algorithmus. Für die Analyse wurden jeweils verschiedene Datensätze aus Sequenzier-Projekten genutzt

3.4 Datenqualität von Sequenzen in Datenbanken

3.4.1 *Die Annotation*

Um jede Sequenz, die in einer Datenbank gespeichert ist, jederzeit eindeutig identifizieren zu können, wird ihr eine „*Accession Number*“[1] zugeordnet. Die drei wichtigsten Datenbanken in denen Sequenzdaten abgelegt sind, sind Genbank, EMBL[2] und DDBJ[3]. Durch die *International Nucleotide Sequence Database Collaboration*[4], werden sie täglich miteinander abgeglichen, sodass identische Datensammlungen vorliegen. Jeder Datenbank-Eintrag ist zusammengesetzt aus einer Annotation, d.h. der Beschreibung der Sequenz, deren Herkunft, Funktion, Verweise auf Literatur usw. und der Sequenz an sich. In Abbildung 3-8 ist ein Ausschnitt eines Eintrags im GenBank-Format dargestellt. [Hansen, 2004]

```
LOCUS       LISOD                   756 bp    DNA     linear   BCT 30-JUN-1993
DEFINITION  Listeria ivanovii sod gene for superoxide dismutase.
ACCESSION   X64011 S78972
VERSION     X64011.1  GI:44010
KEYWORDS    sod gene; superoxide dismutase.
SOURCE      Listeria ivanovii
  ORGANISM  Listeria ivanovii
            Bacteria; Firmicutes; Bacillales;
REFERENCE   1  (bases 1 to 756)
  AUTHORS   Haas,A. and Goebel,W.
  TITLE     Cloning of a superoxide dismutase
            functional complementation in Esc
            of the gene product
  JOURNAL   Mol. Gen. Genet. 231 (2), 313-322
  MEDLINE   92140371
REFERENCE   2  (bases 1 to 756)
  AUTHORS   Kreft,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-APR-1992) J. Kreft,
            Universitaet Wuerzburg, Biozentru
FEATURES             Location/Qualifiers
     source          1..756
                     /organism="Listeria ivanovii"
                     /strain="ATCC 19119"
                     /db_xref="taxon:1638"
                     /mol_type="genomic DNA"
     RBS             95..100
                     /gene="sod"
     gene            95..746
                     /gene="sod"
     CDS             109..717
                     /gene="sod"
                     /EC_number="1.15.1.1"
                     /codon_start=1
                     /transl_table=11
                     /product="superoxide dismutase"
                     /db_xref="GI:44011"
                     /db_xref="GOA:P28763"
                     /db_xref="InterPro:IPR001189"
                     /db_xref="UniProtKB/Swiss-Prot:P28763"
                     /protein_id="CAA45406.1"
                     /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVS
                     GHAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLK
                     AAIESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPV
                     LGLDVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
     terminator      723..746
                     /gene="sod"
ORIGIN
        1 cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat
       61 gtaatttctt .........
//
```

Abbildung 3-8: GenBank Eintrag einer DNA-Sequenz

Ein Datenbank- Eintrag besteht aus einer Annotation mit Angaben zur Sequenzlänge, Definition der Sequenzart, Organismus, Literaturverweise und Autoren (links) sowie genaueren Beschreibungen zu Sequenzabschnitten (z.B. Bereich von Genen) und der Sequenz an sich (rechts).

[1] abgekürzt als AC, besteht aus einem Buchstabe und fünf Zahlen (z.B. D32432) oder zwei Buchstaben und sechs Zahlen (z.B. EF748374)
[2] European Molecular Biology Laboratory
[3] DNA Data Bank of Japan
[4] abgekürzt als INSDC, ist eine Kooperation zwischen GenBank, EMBL und DDBJ

3.4.2 *Sequenzen des INSDC*

Alle Sequenzen, die im INSDC abgelegt werden, können für verschiedenste Analysen herangezogen werden. Durch die Entwicklung des *„Basic local alignment search tools“* [1] kann in der Datenbank nach Sequenzen gesucht werden, die zu einer Eingabesequenz am ähnlichsten sind. Dies gibt Forschern die Möglichkeit Proben zu identifizieren oder taxonomische Zuordnungen einfach und schnell durchzuführen.
Die Qualität der Ergebnisse ist jedoch von verschiedenen Faktoren abhängig. Es muss eine große Anzahl Sequenzen unterschiedlichster Gattungen und Arten in den Datenbanken gespeichert sein um möglichst viele Organismen zu repräsentieren. Die Einträge müssen ausreichend annotiert und dem richtigen Organismus zugeordnet sein und es sollten standardisierte, universelle Verfahren für einzelne Zielstellungen zur Verfügung stehen. Überprüft man diese Kriterien am Reich der Pilze, kann jedoch keins dieser erfüllt werden: Die ITS Region[2] wurde bei weniger als 1 % der geschätzten 1,5 Mio. Arten sequenziert. Des Weiteren kann davon ausgegangen werden, dass ca. 20 % aller Einträge in Bezug auf die Spezies falsch annotiert sind. Außerdem haben Untersuchungen gezeigt, dass ca. 30 % der Pilzsequenzen keiner bestimmten Art zugewiesen wurden. Dies kann bei der Suche nach bestimmten Sequenzen mit BLAST zu falschen oder unzureichenden Ergebnissen führen [Nilsson, 2006].
Betrachtet man die geographische Verteilung der Herkunft der Sequenzen, stammen die meisten Einträge aus Nord-Amerika, Europa, China und Japan. Dem gegenüber stehen große Gebiete wie Afrika, von denen keine oder nur sehr wenige Einträge in der Datenbank zu finden sind, was sich nachteilig auf Untersuchungen der Sequenzvariabilität geographisch getrennter Populationen auswirkt. Ein weiteres Problem bei der Suche von Pilzen-Sequenzen ergibt sich durch die Vielzahl an möglichen Synonymen, unter denen Pilzarten bezeichnet werden können. Ohne vorhandenes Fachwissen könnten so Einträge übersehen oder als nicht wichtig angesehen werden. Eine mögliche Lösung des Problems stellen spezielle Pilz-Datenbanken[3] dar, in denen mögliche Synonyme einer Art aufgelistet werden [Ryberg, 2009].

[1] abgekürzt als BLAST [Altschul, 1990]
[2] Die ITS Region ist ein Teil der ribosomalen DNA und wird als Standardregion für Spezies- Identifikationen bei Pilzen verwendet.
[3] zum Beispiel die MycoBank (http://www.mycobank.org)

3.4.3 *Spezielle Datenbanken*

Das Problem der ungenauen oder falschen Annotation ist Auslöser dafür, Datenbanken oder Anwendungen zu entwickeln, durch die qualitativ hochwertige Sequenzen oder Einträge gefiltert werden können. Beispielhaft ist hier die *fyBase* der Firma *Nadicom* zu nennen. Es handelt sich hierbei um eine kommerzielle Datenbank, in der mehr als 30.000 ITS Sequenzen abgespeichert sind und zur Identifikation von Pilzen und Hefen genutzt werden kann. Des Weiteren liefert *fyBase* Metadaten zu den Organismen wie morphologische Daten, Stoffwechsel-Eigenschaften oder taxonomische Einordnung [Nüßlein, 2009]. Ausserdem ist die öffentlich zugängliche Datenbank *ArbSilva* zu nennen, in welcher ausschließlich ribosomale DNA –Sequenzen von Bakterien, Archaen und Eukaryoten gespeichert sind. Diese werden nur in die Datenbank aufgenommen wenn die Sequenzqualität über einem bestimmten Schwellwert liegt [Pruesse, 2007]. Eine weitere Quelle für qualitativ hochwertige Sequenzen von Pilzen ist die *Barcode of Life Database*[1] (vgl. Abbildung 3-9). Eine nähere Beschreibung des Sequenz-Barcodes sowie deren Bestimmung ist in Kapitel 4 zu finden.

Abbildung 3-9: Startseite der Barcode of Life Database

In der *Barcode of Life Database* werden nur qualitativ hochwertige DNA-Sequenzen aufgenommen. Diese werden als Barcodes bezeichnet.

[1] Abgekürzt als BOLD

4 Barcoding

In der Taxonomie[1] werden vor allem morphologische Merkmale, wie Form, Größe oder Farbe einzelner Körperteile, benutzt um Arten eindeutig identifizieren zu können [CBOL, 2011], [Stoeckle, 2008]. Die bis heute mit dieser Methode beschriebenen ca. 1,8 Mio. Spezies stellen jedoch nur einen Bruchteil der Gesamtzahl an Organismen dar. Schätzungen zeigen auf, dass 10 – 15 Millionen, [May, 2009], [Hebert, 2003] nach neueren Ergebnissen möglicherweise auch 100 Millionen Arten auf unserem Planeten leben [Steinke, 2006]. Wenn davon ausgegangen wird, dass einige Taxonomen ca. 1000 Organismen unter zu Hilfenahme verschiedener Literaturvergleiche und Sammlungen identifizieren kann, benötigt man im günstigsten Fall 10.000 bis 15.000 Experten um die geschätzte Anzahl von Arten abzudecken [May, 2009], [Hebert, 2003]. Da aber bereits heute die Anzahl der zur Verfügung stehenden Taxonomen immer mehr schwindet, ist es schwer vorstellbar, dass diese Zahlen erreicht werden. Es wird eine neuere Technologie benötigt, die die Identifizierung und Beschreibung von Arten vereinfacht und unterstützt, sodass auch Laien vielversprechende Ergebnisse hervorbringen können und die Experten sich auf die Entdeckung neuer Arten konzentrieren.

4.1 Grenzen der morphologischen Identifizierung

Die bisher verwendeten morphologischen oder auch verhaltensbiologischen Eigenschaften von Organismen führen, bedingt durch verschiedene Faktoren, nicht immer zu einem eindeutigen und richtigen identifiziertem Ergebnis.
Phänotypische Plastizität und genetische Variationen können zu falschen Ergebnissen bei der Identifikation führe [Schmidt, 2001]. Zum Beispiel verändern einige Molcharten ihre körperlichen Eigenschaften, wie schwärzere Schwanzflossen oder stärkere Schwanzmuskeln, wenn sie unter Anwesenheit von Libellenlarven aufwachsen, um ihre Überlebenschance zu steigern [Schmidt, 2001]. Des Weiteren gibt es einige Taxa, bei denen die morphologischen Abstufungen so gering sind, dass sie nicht in einzelne Arten unterschieden werden können [Hebert, 2003]. Der Dickkopffalter *Astraptes fulgerator* besitzt verschiedene Muster im Larvenstadium. Die erwachsenen Schmetterlinge besitzen jedoch keine morphologischen Unterschiede mehr. Erst mit Hilfe der

[1] Beschäftigt sich mit der Identität von Organismen sowie deren Beziehungen zueinander [Savolainen, 2005]. Ein Wissenschaftler, der auf diesem Gebiet tätig ist, wird Taxonom genannt.

Barcode Technologie in Verbindung mit dem Fressverhalten und Muster der Raupen konnte geklärt werden, dass es sich um verschiedene Arten (insgesamt 10) handelt [Hebert, 2004], [Stoeckle, 2008].

Abbildung 4-1: Morphologische Unterschiede bei den Raupen von A. fulgerator

Mit Hilfe der Barcode Technologie konnte nachgewiesen werden, dass der Dickkopffalter *A. fulgerator* zehn verschiedene Arten in sich vereint.

Die zur eindeutigen Identifikation beschriebenen morphologischen Eigenschaften sind meist nur auf ein Geschlecht oder eine Entwicklungsstufe anwendbar. Eine Art, die verschiedene Entwicklungsstufen durchläuft, ist dadurch schwerer bis unmöglich zu bestimmen [Hebert, 2003]. Am Beispiel des Dickkopffalters besitzen die Raupen unterschiedliche Musterungen (vgl. Abbildung 4-1) und zeigen artspezifisches Fressverhalten, welches sie unterscheidbar macht. Die adulte Form besitzt allerdings keine Abgrenzungen zwischen den Arten [Stoeckle, 2008]. Für die Bestimmung der Spezies sind auch minimalste Unterschiede von Bedeutung. Diese zu erkennen bleibt meist die Aufgabe von Experten, da diese das nötige Fachwissen und Erfahrung besitzen. Dieser enorme Anspruch an Genauigkeit und Sachverstand führt jedoch unausweichlich auch zu Fehlinterpretationen [Hebert, 2003].

Anhand dieses Beispiel wird deutlich, dass das Barcoding ein wichtiges Tool für die bisher angewandten taxonomischen Arbeiten darstellt.

4.2 Der DNA - Barcode

Ähnlich den Barcodes („Strichcodes“) im Supermarkt, genannt UPC (universal product code), besitzt das Genom jeder Spezies eine spezifische und einzigartige Abfolge von Basen (Nukleotiden). Somit kann eine DNA-Sequenz, die aus einer unbekannten Probe isoliert wurde, mit Hilfe von Sequenzvergleichen einer Referenz-Datenbank zur schnellen, eindeutigen und möglicherweise automatisierbaren Speziesidentifikation dienen [Hebert, 2005]. Aus dieser Idee, die erstmals durch Paul D. N. Hebert im Jahr 2003 verbreitet wurde [Hebert, 2003], entstand das DNA-Barcoding [Stoeckle, 2003]. Ein DNA-Barcode wird dabei als kurze, standardisierte Genomregion bezeichnet [Hebert, 2005]. Er soll als Hilfsmittel für taxonomische Arbeiten eingesetzt werden um diese zu verbessern und zu erweitern und die Identifizierung von Arten zu erleichtern [Stoeckle, 2003]. Er sollte nicht dazu verwendet werden, unbekannte Arten in den Tree of Life einzuordnen oder phylogenetische Analysen zu ersetzen [Moritz, 2004].

4.2.1 *Anforderungen an einen DNA-Barcode*

Da ein Barcode zur eindeutigen Art-Identifizierung von Organismen dienen soll, sollte er idealerweise signifikante genetische Variabilitäten und Divergenzen zwischen verschiedenen Arten (interspezifisch) [Kress, 2008] aufweisen. Diese interspezifischen Variationen müssen sehr viel höher als die intraspezifischen[1] Unterschiede sein (vgl. Abbildung 4-2) [Savolainen, 2005], [Hajibabaei, 2007] und wird auch als Barcode-Gap bezeichnet [Meyer, 2005]. Für die Entwicklung universeller PCR-Primer sollte die Barcode-Sequenz angrenzende konservierte Bereiche aufweisen, um möglichst viele Barcodes aus den verschiedensten Gattungen schnell extrahieren zu können [Kress, 2008]. Des Weiteren ist es von Vorteil die komplette Barcode-Sequenz mit einer einzigen Sequenzierungsreaktion zu bestimmen, um eine Assemblierung und damit möglichen Fehlerquellen zu umgehen [Kress, 2008]. Dadurch ist die mittlere Länge eines Barcodes auf ca. 400-800 Nukleotide begrenzt (abhängig von der gewählten Sequenzier-Methode) [Kress, 2008]. Außerdem nimmt die Anzahl der möglichen diagnostischen Unterschiede (variierende Basen) zwischen zwei Sequenzen Einfluss auf die Länge des Barcodes. Theoretisch würde eine DNA-Sequenz mit der Länge von 15 Nukleotiden (ergibt $4^{15} \cong 1$ Milliarde verschiedene Kombinationen der vier Basen Adenin, Cytosin, Guanin und Thymin) ausreichen, um die geschätzten 10 – 100 Mio.

[1] innerhalb einer Art

Arten auf der Erde eindeutig differenzieren zu können. Jedoch ist nicht jede Base mit der gleichen Wahrscheinlichkeit austauschbar. Aufgrund funktioneller Unterschiede bestimmter DNA-Abschnitte (codierende/nicht codierende Bereiche) können Basen konserviert oder hoch variabel sein [Hebert, 2003].

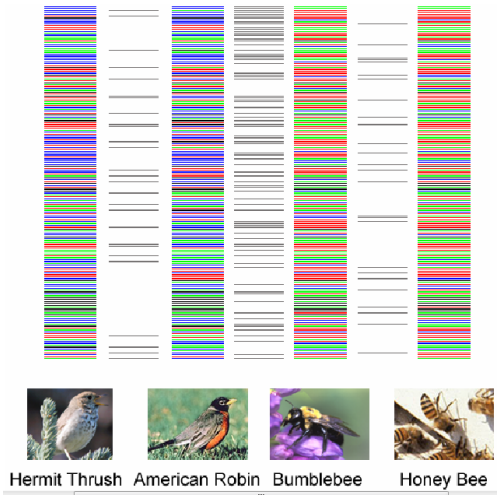

Abbildung 4-2: Beispiele für Barcodes und deren diagnostischen Unterschiede

Die beiden Vogelarten (links) und die Bienenarten (rechts) zeigen jeweils geringere Unterschiede in ihren Barcodes als zwischen Vogelart und Bienenart. Die Abfolge der Nukleotide wird als farbcodiertes Bild wiedergegeben [nach Stoeckle, 2004].

Durch Insertionen, Deletionen oder Substitutionen wird die genetische Information an einzelnen Positionen verändert und kann somit größere intraspezifische Unterschiede verursachen [Merkl, 2009 a]. Zum Beispiel bei Populationen einer Art mit verschiedenen geographischen Standorten (geographische Isolation) [Meyer, 2005]. Um die Auswirkung dieser Effekte möglichst gering zu halten, liegt der Fokus für mögliche Barcode-Regionen auf proteincodierenden Genen [Hebert, 2003].

4.2.2 *Anwendungsbeispiele*

Die Verwendung des Barcodes ist in einer Vielzahl von Lebensbereichen und wissenschaftlichen Untersuchungen möglich. Zum Beispiel können Biologen in Feldstudien schnell und zuverlässig die Biodiversität eines Biotops analysieren. Gesundheitsorganisationen können gezielter Kontrollen krankheitsübertragender Insekten durchführen oder Ärzte schneller mögliche Infektionskrankheiten erkennen. Im Lebensmittelbereich können Waren auf die Richtigkeit ihrer Etikettierung überprüft werden oder Schädlinge auf Feldern identifiziert und schneller Gegenmaßnahmen ergriffen werden [Stoeckle, 2008].

4.3 Barcoderegionen

4.3.1 *Cytochromoxidase Untereinheit I*

Die mitochondriale Cytochromoxidase Untereinheit I[1] ist ein Protein, welches aus zwölf Transmembrandomänen[2] besteht (vgl. Abbildung 4-3 und Abbildung 4-4). Die höchste Variabilität findet man in den entsprechenden Genabschnitten, welche für die Domänen 1, 3 und 4 sowie dem internen Loop I4 codieren. Dabei substituieren vor allem die dritten Positionen eines Codons[3], aufgrund der redundanten Aminosäurecodierung. [Steinke, 2006] Daraus ergibt sich eine Barcode-Sequenz mit einer Länge von 648 bp am 5´- Ende des COI-Gens. Dies entspricht dem N-Terminalen Ende des Proteins und ist in Abbildung 4-3 als hellblauer Bereich hervorgehoben. Die orange markierten Teile des Proteins stellen stark konservierte Primer-Regionen dar, mit deren Hilfe die Barcode-Sequenz extrahiert werden kann [Hajibabaei, 2005].

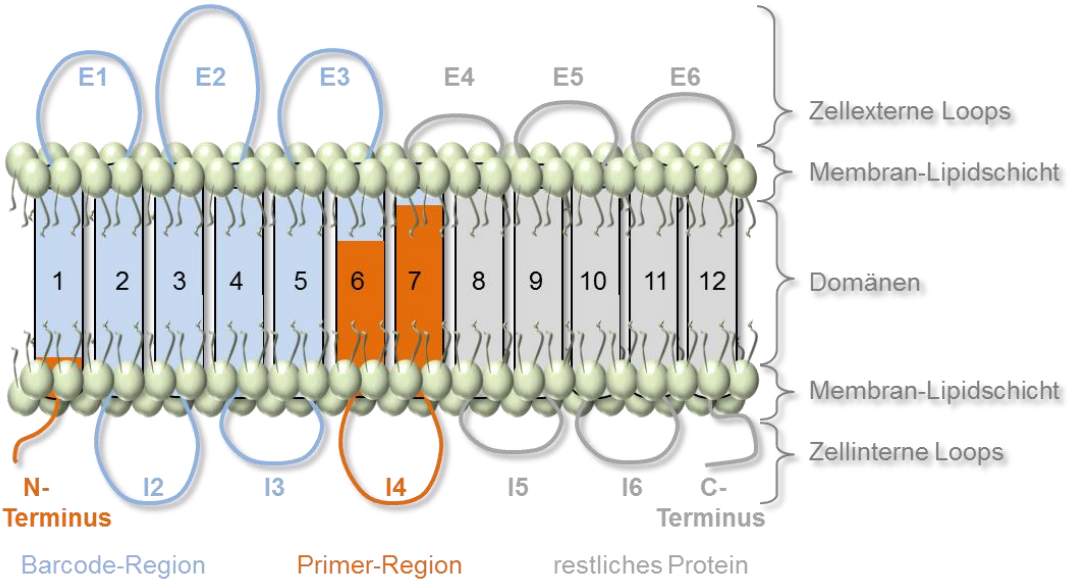

Abbildung 4-3: Cytochromoxidase I Barcode-Region

Dargestellt ist die Sekundärstruktur von COI in einer Biomembran. Der blau markierte Bereich des Proteins entspricht der Barcode-Region des entsprechenden Gens. Die orange markierten Bereiche geben die flankierenden Primer-Regionen wieder. Besonders variabel sind die Domänen 1, 3 und 4 sowie der Loop I4

[1] Abgekürzt als COI oder COX1
[2] Transmembrandomänen sind Proteinketten, die eine Biomembran durchspannen.
[3] Jedes Protein besteht aus aneinander geketteten Aminosäuren, welche jeweils durch drei Nukleotiden in der DNA, den Codons, codiert werden. Für manche Aminosäuren gibt es mehrere mögliche Codons, wodurch Substitutionen in der DNA möglich sind ohne Auswirkung auf das Protein.

Die Verwendbarkeit dieser Sequenz als Barcode wurde bereits an mehreren Tiergruppen, wie zum Beispiel bei Fischen, Vögel, Springschwänzen, Spinnen oder Motten, belegt [Hajibabaei, 2005]. Dabei konnten bei eng verwandten Arten ausreichend diagnostische Unterschiede zwischen den jeweiligen COI-Sequenzen festgestellt werden (ca. 50 Substitutionen in 500 bp) [Hebert et al. 2003b]. Intraspezifisch blieb die Variation jedoch sehr gering (2 Substitutionen in 100 bp) [Hebert et al. 2003a], [Stoeckle, 2003].

Die Wahl einer mitochondrialen DNA-Sequenz als Barcode bringt außerdem viele Vorteile mit sich: Es existieren keine Introns und Rekombinationen können, aufgrund der mütterlichen Vererbung nicht stattfinden. [Steinke, 2006] Des Weiteren gibt es schon eine Reihe etablierter Primer und die Sequenz kommt in der Zelle, im Gegensatz zum Nukleus, relativ häufig vor. Dadurch wird die Extraktion vereinfacht [Stoeckle, 2008].

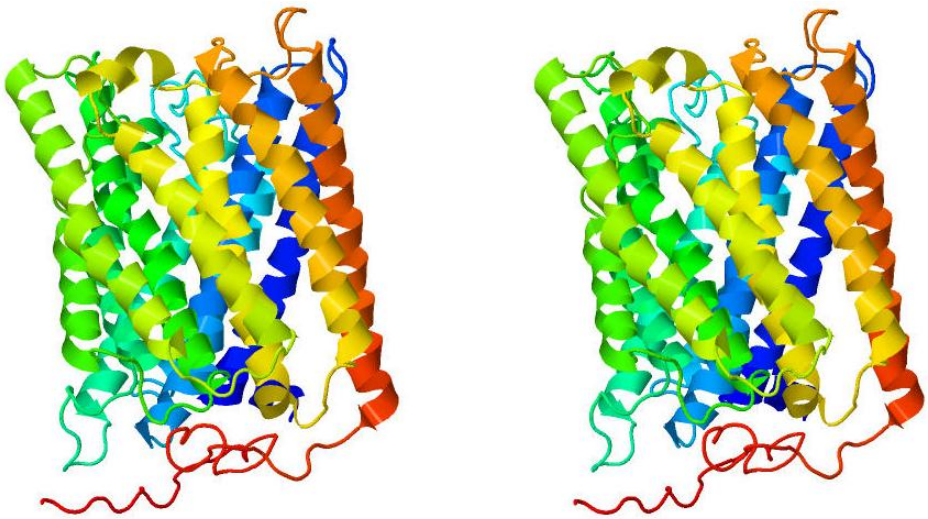

Abbildung 4-4: Cytochromoxidase Untereinheit I (COI)

Das 5'-Ende des COI-Gens ist eine anerkannte, ca. 648 bp lange, Barcode-Region für Tiere. Die Abbildung ist eine stereoskopische (Kreuzblick) Darstellung der Tertiärstruktur der mitochondrialen Cytochromoxidase Untereinheit I des Rinderherzens (PDB Eintrag 2OCC).

4.3.2 *Ribosomale DNA*

Aufgrund der möglicherweise vorhandenen Introns in der mitochondrialen DNA bei Pilzen könnte die Gewinnung eines COI-Barcodes erschwert werden [Begerow, 2010]. Dies veranlasste die Suche nach alternativen Barcode-Sequenzen. Eine Möglichkeit stellt die nukleare ribosomale DNA dar (vgl. Abbildung 4-5). Sie besteht aus drei relativ gut konservierten, für Teile des Ribosom codierenden, Abschnitten (LSU, SSU und

5,8S) sowie zwei variablen „*internal transcribed spacer*“ [1]-Bereichen. Die mögliche Verwendung der ITS-Region als Barcode, wird dadurch bekräftig, dass sie schon seit einigen Jahren für Pilzidentifikationen verwendet wird [Begerow, 2010].

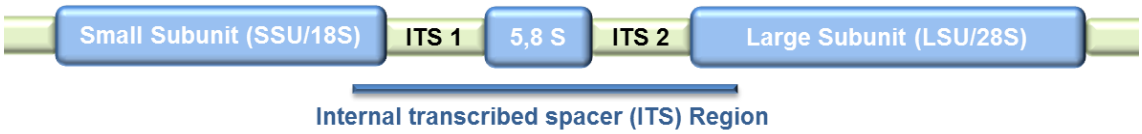

Abbildung 4-5: Aufbau der ribosomalen DNA

Die ribosomale DNA besteht aus fünf Komponenten: Drei, für Teile des Ribosoms, codierende Gene (die kleine Untereinheit (SSU oder 18S), die große Untereinheit (LSU oder 28S) und die 5,8S Einheit) sowie zwei Internal Trascribes Spacer Regionen (ITS1 und ITS2). Der Sequenzabschnitt ITS1-5,8S-ITS2 wird auch als ITS-Region bezeichnet.

Jedoch wurden auch Argumente gegen die Verwendung als Barcode geäußert. Zum Beispiel gibt es bei manchen Gattungen eine zu hohe intraspezifische Variabilität [Begerow, 2010] und die Sekundärstruktur der transkribierten ribosomalen RNA ermöglicht Insertionen und Deletionen [Strinke, 2006]. Des Weiteren ist die ITS-Region nur ca. 500 bp lang, was zu Problemen bei der Bestimmung von artenreichen Gattungen führen würde (nicht genug diagnostische Unterschiede) [Seifert, 2009]:

4.3.3 *Barcodes für Pflanzen*

In Pflanzen entwickelte sich das COI-Gen langsamer als bei Tieren und enthält dadurch weniger Variationen in der genetischen Information. Aus diesem Grund wurde nach alternativen Barcode-Regionen gesucht. Nach einigen Studien erklärte die *CBOL Plant Working Group* im Jahr 2009 die Kombination von *matK, rbcL* und *trnH-psbA* als Barcode für Pflanzen [Janzen, 2009], [Kress, 2009]. Dabei handelt es sich bei *matK* und *rbcL* um Proteine codierende DNA-Sequenzen und bei *trnH-psbA* um nicht-codierende Regionen, welche in der Plastid-DNA[2] lokalisiert sind [Chase, 2009].

[1] abgekürzt als ITS
[2] Plastide sind Organelle in der Pflanzenzelle, auch als Chloroplasten bezeichnet

4.4 Die Barcoding Pipeline

Alle Barcode Projekte besitzen den gleichen Ablauf, um Organismen einen standardisierten Barcode zuweisen zu können. Diese BARCODING PIPELINE ist in Abbildung 4-6 dargestellt. Der erste Schritt stellt das Sammeln von Proben dar. Als Quelle können Museen, Herbarien, Zoos, Aquarien, eingefrorene Gewebeproben, Samenbanken oder auch Organismenproben direkt aus der Umwelt dienen. Aus diesen Materialen wird Gewebe entnommen. Im zweiten Schritt werden durch laboranalytische Verfahren die Ziel-DNA extrahiert, gereinigt und die entsprechende Sequenz mittels Sequenzierung (vgl. Kapitel 2) bestimmt. Für die einheitliche Ausführung der Laborarbeiten und der Vermeidung von Fehlern wurden verschiedene Protokolle [Ivanova, 2011] von dem *Consortium for the Barcode of Life*[1] erarbeitet und online zur Verfügung gestellt. Das größte Barcoding-Projekt ist das *International Barcode of Life Project*[2] , welches durch das *Biodiversity Institute of Ontario at the University of Guelph* organisiert wird. Ihr Ziel ist es, in fünf Jahren fünf Millionen Barcodes von 500.000 Arten zu produzieren. Die gewonnen Barcodes werden dann, im dritten Schritt, an Datenbanken übermittelt und dort verwaltet, analysiert und validiert. Zum Beispiel ist das *Barcode of Life Data Portal* [3] eine Datenbank, in der vor allem Identifikationen, Datensatzanalysen und Visualisierungen der Ergebnisse von Sequenzen durchgeführt werden können. [BOL Data Portal] Nachdem eine Sequenz hinsichtlich ihrer Sequenzqualität geprüft, Verlinkungen zur Spezies und Entnahmestelle der Probe gemacht sowie die Rohdaten der Sequenzierung zur Verfügung gestellt wurden, kann ein öffentlich zugänglicher Eintrag in eine Datenbank erfolgen. Das Hauptziel ist dabei ein schneller und unkomplizierter Datenzugang um unbekannte Proben zu einer Art zuzuordnen. Dies wird zum einen in den Datenbanken des INSDC, die den Datenstandard für Barcode Sequenzen des CBOL akzeptiert haben [Hanner, 2009], und zum anderen im *Barcode of Life Data System*[4] realisiert. In der BOLD hat man zusätzlich die Möglichkeit bereits verifizierte Barcode-Daten zu sammeln, verwalten und zu analysieren [Ratnasingham, 2007]. Der letzte Schritt der Barcoding-Pipeline stellt die Identifizierung von Proben dar. Das beste Alignment[5] zwischen Query- und Referenz-Sequenz ist dabei ausschlaggebend [CBOL, 2011].

---

[1] abgekürzt als CBOL, fördert Barcoding, produziert jedoch keine eigenen Barcode-Sequenzen.
[2] abgekürzt als iBOL
[3] abgekürzt als BOL Data Portal, verfügbar unter http://bol.uvm.edu/index.php
[4] abgekürzt als BOLD
[5] ist ein Vergleich zwischen zwei oder mehr DNA- oder Proteinsequenzen um mögliche Unterschiede zu erkennen und Ähnlichkeiten zu detektieren

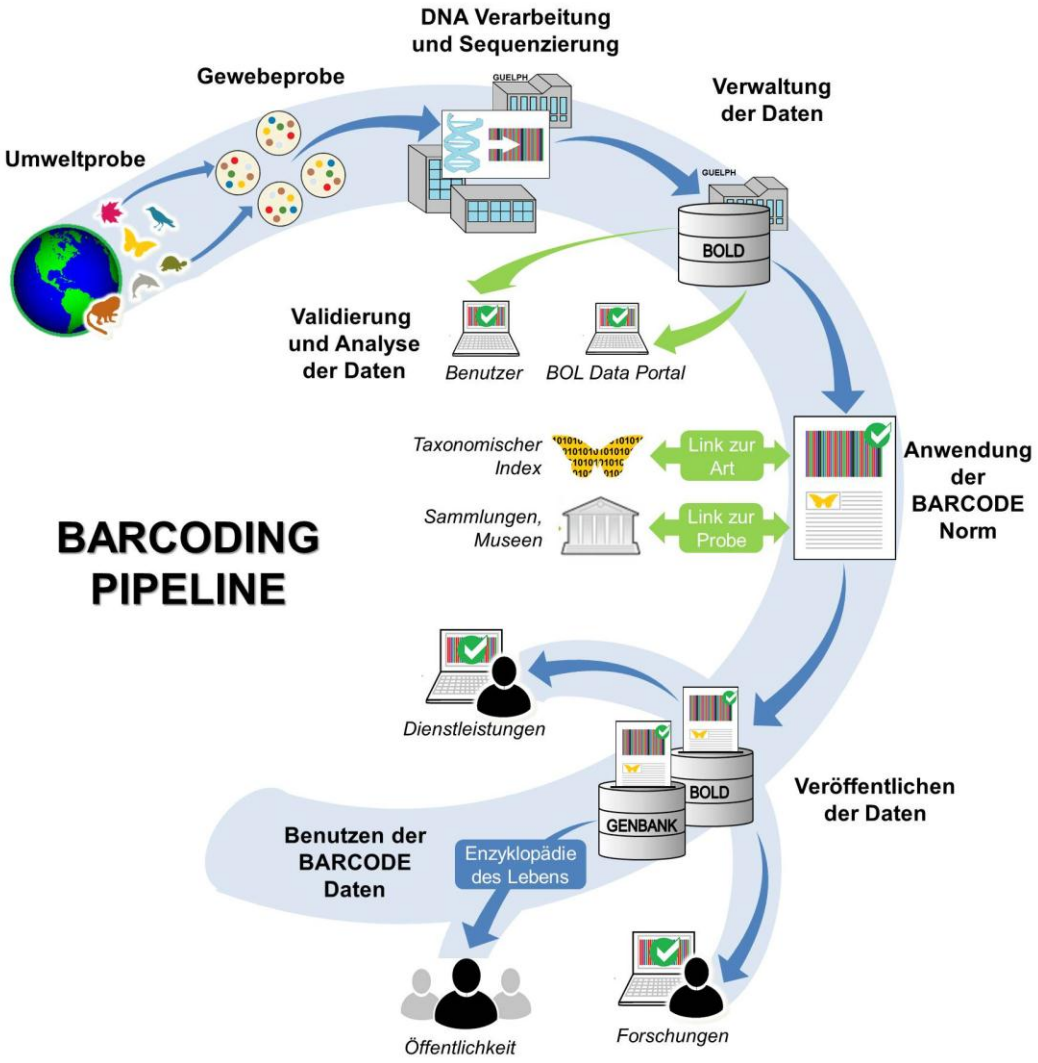

Abbildung 4-6: Barcoding Pipeline

Die Barcoding-Pipeline beschreibt die notwendigen Schritte um zum Beispiel aus einer Umwelt-Probe Barcodes herzustellen. Dafür muss die DNA aus den Proben extrahiert und anschließend sequenziert werden. Nach der Übermittlung der Daten an eine Datenbank erfolgt die Analysierung und Verifizierung der Sequenzen. Erfüllt die Sequenz den Barcode-Datenstandard wird sie mit verschiedenen Verlinkungen zu z.B. Spezies oder Entnahmestelle der Probe ausgestattet und in einer entsprechenden Datenbank veröffentlicht. Diese Sequenzen können dann zur Spezies-Identifikation in verschiedensten Anwendungs-gebieten dienen.

4.5 Barcode of Life Data System

Das *Barcode of Life Data System (BOLD)* ist eine Datenbank in der Barcodes erfasst, gespeichert, analysiert und publiziert werden [Ratnasingham, 2007].
Ursprünglich wurde diese Datenbank im *Biodiversity Institute of Ontario* der *University of Guelph* mit der Idee, dass Barcodes mit Information der Probeorganismen und zusätzlichen Daten verlinkt werden sollten, entwickelt. Des Weiteren muss die durch Hochdurchsatzverfahren schnell ansteigende Datenmenge organisiert werden [Hajibabaei, 2005]. BOLD ist aus drei Einheiten zusammengesetzt: dem *External Connectivity System (ECS)*, dem M*anagement And Analysis System (MAS)* und einer *Species Identification Engine*.
Da die Datenbank frei zugänglich ist, können Forschungsgruppen leicht ihre eigenen Barcode-Daten veröffentlichen. Dafür wird ein Datenstandard vorausgesetzt. Um einen Eintrag, der eine ID-Nummer der Probe sowie eine taxonomische Zuordnung enthalten muss, als Barcode bezeichnen zu können, werden folgende Angaben benötigt: Der Name der Spezies, Referenz Daten (Katalog Nummer und Lagerort), Angaben zur Probe (Sammler, Datum und Ort der Probenahme mit GPS Koordinaten), eine Bezeichnung der Probe, COI –Sequenzen müssen eine Mindestlänge von 500 bp aufweisen, die Namen der PCR Primer und deren Sequenz sowie alle EPG.
Daten die diesen Standard nicht erfüllen, können trotzdem in der Datenbank vorhanden sein, dürfen jedoch nicht als Barcode bezeichnet werden. Des Weiteren werden von BOLD die Daten geprüft und falls mögliche Fehler entdeckt werden, wird der Eintrag entsprechend gekennzeichnet. Die Bewertung von EPG wird mit Hilfe des PHRED Algorithmus durchgeführt, wobei jede Sequenz in eine von vier Kategorien eingeteilt (vgl. Tabelle 4-1) wird [Ratnasingham, 2007].

Tabelle 4-1: Einteilung der Qualität von EPG in BOLD

In BOLD wird jedes Elektropherogramm in eines der vier Qualitätsstufen eingeteilt. Die Bewertung wird anhand des Durchschnitts aller PHRED-Werte einer Sequenz durchgeführt.

Einstufung in BOLD	Kriterium
Failed	keine Sequenz vom EPG ableitbar
Low quality	PHRED Werte < 30
Medium quality	PHRED Werte 30 – 40
High quality	PHRED Werte > 40

Die Daten werden als zweiphasiges System abgespeichert. Es gibt eine Proben-Seite und eine Sequenz-Seite. Diese sind in Abbildung 4-7 dargestellt.

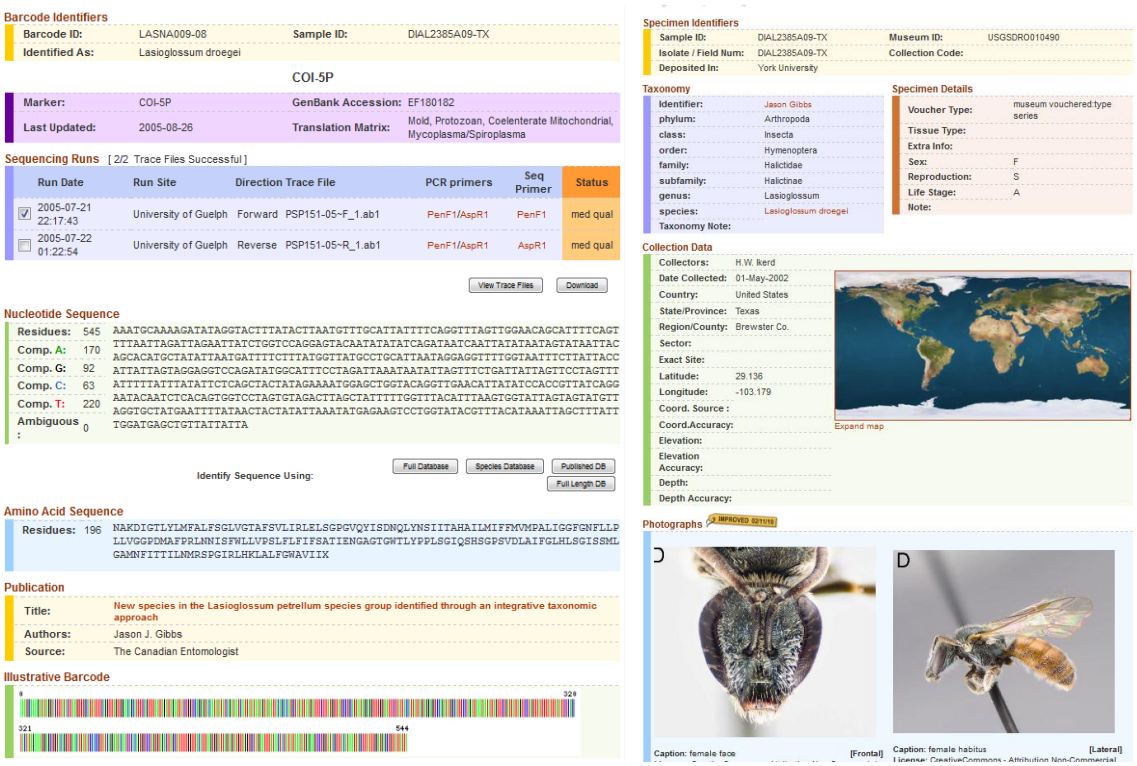

Abbildung 4-7: Abspeicherung und Darstellung der Daten in BOLD

Für einen Eintrag können zwei verschiedene Seiten aufgerufen werden. Die eine enthält alle Sequenzdaten eines Eintrags (links) und die andere Daten zu Probe, geographische Angaben sowie Bilder (rechts).

Diese Trennung der verschiedenen Daten wird auch bei der Download-Funktion beibehalten. Man kann entweder die FASTA[1]-Files ausgewählter Sequenzen, Spreadsheets (Excel-Tabellen), in denen alle wichtigen Details zu Einträgen enthalten sind, oder Trace-Files herunterladen [Ratnasingham, 2007].

Es stehen auch einige analytische Tools, wie zum Beispiel die Analyse von EPGs deren Darstellung sowie Assemblierungen verschiedener Sequenzen. *Im Identification System* können zudem unbekannte Sequenzen gegen die Datenbank abgeglichen werden und so eventuell taxonomisch eingeordnet werden [Ratnasingham, 2007].

[1] FASTA ist eine einfache Textdatei, deren Inhalt aus zwei Komponenten aufgebaut ist. Die erste Zeile beginnt mit einem „>“ und enthält Informationen über die am der zweiten Zeile dargestellten Sequenz. Dies können zum Beispiel ID-Nummern, Organismennamen oder Genbezeichnungen sein.

5 Sequenzanalyse: Eine Strategie

Das Erstellen von Strategien hat bei umfangreichen Projekten einen großen Stellenwert, da Arbeitsschritte leicht überprüft bzw. korrigiert und fremde Personen leicht in die Prozessstruktur eingewiesen werden können. Die hier vorgestellte Strategie stellt alle Arbeitsschritte, die für die erfolgreiche Umsetzung dieses Projektes notwendig sind, zusammengefasst dar. Sie ist ein Leitfaden für die Bearbeitung und Analyse von DNA-Sequenzen und zudem die Grundlage für die Entwicklung einer Software mit deren Hilfe auch andere biologische Fragestellungen gelöst werden können.

5.1 Überblick

Die Strategie kann in fünf verschiedene Schritte eingeteilt werden. Diese sind in Abbildung 5-1 dargestellt. Zu Beginn wird eine Sequenzbibliothek erstellt, in der alle möglichen Informationen die man benötigt, abgespeichert sind. Im Zweiten Schritt wird überprüft, ob einige Sequenzen als fragmentiert, d. h. die eigentliche Zielsequenz besteht aus vielen einzelnen Sequenzen, vorliegen. Um diese zusammenzufügen wird eine Assemblierung durchgeführt. Die Gruppierung von Sequenzen nach ihren Ähnlichkeiten erfolgt durch ein multiples Sequenz-Alignments. Diese geben Aufschluss über verwandtschaftliche Beziehungen von Organismen oder helfen dabei, konservierte und variable Bereiche in Sequenzen zu finden. Der vierte Schritt beinhaltet die Bestimmung von Primer und/oder Sonden für Amplifizierungen, Sequenzierungen (mit Hilfe der Primer) oder Identifizierungen (mit Hilfe von Sonden) durchführen zu können. Der letzte Schritt beinhaltet die Entwicklung eines Test-Kits. Dabei kann es sich je nach Aufgabenstellung um Laborprotokolle, einfache Teststreifen oder komplexe Microarray-Systeme handeln.

Abbildung 5-1: Strategie Übersicht

Die Strategie ist in fünf Punkten gegliedert. Je nach Aufgabenstellung unterscheiden sich die durchzuführenden Analysen.

5.2 Sequenzbibliothek erstellen

Der erste und wichtigste Schritt ist die Erstellung einer Sequenzbibliothek, in der alle benötigten Informationen abgespeichert werden (vgl. Abbildung 5-2). Für die Gewinnung von Sequenzen gibt es zwei mögliche Datenquellen. Zum einen sind in Datenbanken Millionen von Sequenzen abgespeichert und frei zugänglich. Zum anderen dienen Sequenzierungen dazu, neue Sequenzen zu erhalten. Der Vorteil bei der Benutzung einer Sequenzierung ist, dass man Zugang zu den Rohdaten (EPGs) besitzt.

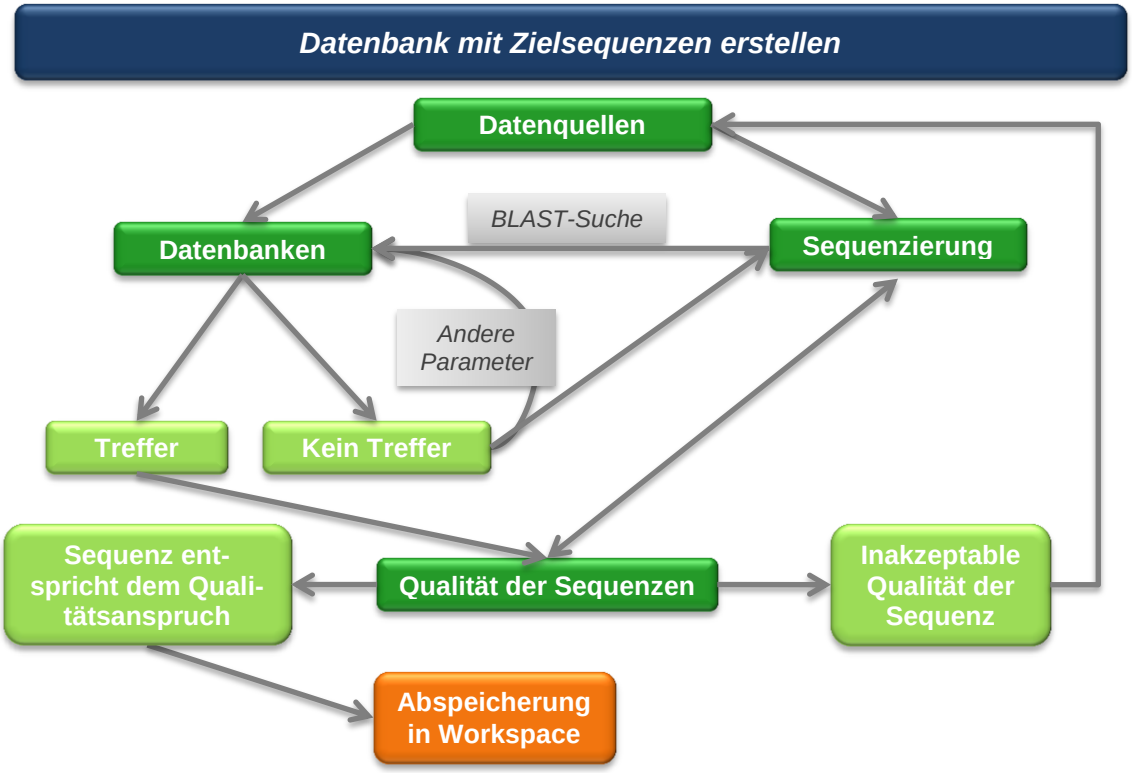

Abbildung 5-2: Erster Schritt - Erstellen einer Sequenzbibliothek

Der erste Schritt der Strategie beinhaltet die Erstellung einer Datensammlung mit allen wichtigen Informationen, die für die gewünschte Analyse notwendig sind. Ein wichtiger Punkt hierbei ist die Bewertung der Sequenzqualität um möglichen Fehlerquellen im weiteren Verlauf vorzubeugen.

In Datenbanken kann man durch zwei verschiedene Verfahren zu Daten gelangen: Mit einer Stichwortsuche oder einer BLAST-Suche. Die Stichwortsuche beinhaltet meist den Namen des gewünschten Organismus sowie die Bezeichnung der gesuchten Sequenz, wie zum Beispiel „*Aspergillus fumigatus,* ribosomal DNA“. Optional können auch logische Operatoren wie UND bzw. ODER verwendet werden um die Ergebnisse besser zu filtern. Meist wird für solche Zwecke eine Datenbank des INDSC verwendet.

Möchte man mit Hilfe einer schon bekannten Sequenz „Query“ oder Eingabesequenz in der Datenbank nach ähnlichen Einträgen suchen, nutzt man zum Beispiel das Tool BLAST. Abhängig von den vorgenommenen Einstellungen werden mit Hilfe des heuristischen Algorithmus[1] mögliche, zur Query homologe[2], Sequenzen in der Datenbank gesucht. Zur Bewertung der statistischen Signifikanz[3] der Ergebnisse wird der E-Wert berechnet. Je kleiner dieser ist (minimal 0) desto höher ist die Wahrscheinlichkeit, dass der Treffer nicht zufällig entstanden ist [Hansen, 2004]. Die Suche in Datenbanken hat den Vorteil, dass man schnell und kostenfrei an eine große Anzahl von Sequenzen gelangt. Nachteil ist jedoch, dass die Suchanfragen stark vom Datenbestand abhängig sind, man oft keinen Zugang zu Rohdaten hat und die Daten nicht immer richtig sind (vgl. Kap. 3.4). Werden keine Sequenzen in den Datenbanken gefunden, obwohl die Suchparameter variiert und Fehlerquellen ausgeschlossen wurden, ist eine Sequenzierung durchzuführen und die Datenqualität zu überprüfen. Sind die Sequenzen als qualitativ hochwertig eingestuft, werden sie lokal gespeichert und sind für weitere Analysen benutzbar. Sollten sie den Anforderungen nicht entsprechen, ist eine erneute Sequenzsuche bzw. Sequenzierung notwendig.

[1] Heuristische Algorithmen werden eingesetzt um große Datenmengen in kürzester Zeit zu durchsuchen. Sie sind jedoch nicht so sensitiv wie dynamische Algorithmen.

[2] Die Ähnlichkeit zwischen zwei Sequenzen ist definiert durch die Anzahl der identischen und substituierten Positionen. Durch die Untersuchung von Sequenzidentität und – Ähnlichkeit kann man Aussagen über die Homologie dieser Sequenzen treffen.

[3] Eine Datenbanksequenz die Ähnlichkeiten zur Query aufweist kann auch zufällig gefunden worden sein. Diese Sequenzen müssen aber nicht unbedingt homolog sein.

5.3 Assemblierung

Da mit einer Sequenzierung maximale Sequenzlängen von 600 – 800 nt erreicht werden können und in Datenbanken oft auch nur Teilsequenzen vorhanden sind, müssen Assemblierungen, d. h. zusammensetzen von Teilsequenzen, durchgeführt werden. Dies wird vor allem bei Genomprojekten angewandt. Dabei müssen mehrere tausend Sequenzen zu größeren so genannten „*contigs*“ (auch als Konsensus-Sequenz bezeichnet) zusammengefasst werden. Das Ziel ist, eine einzige Sequenz zu erhalten, die das Genom eines Organismus repräsentiert. Das Prinzip der Assemblierung ist einfach: In den Ausgangssequenzen werden überlappende Bereiche gesucht und anhand verschiedener Parameter, wie minimale Länge des überlappenden Bereiches, Anzahl identischer Positionen in diesem Bereich oder minimalen Score, überprüft ob diese zwei Sequenzen zusammengefügt werden können oder nicht (vgl. Abbildung 5-3).

Abbildung 5-3. Prinzip der Assemblierung

Eine Assemblierung wird benötigt um mehrere Sequenzen anhand ihrer Ähnlichkeiten zusammenfügen zu können. Diese Methode wird zum Beispiel benötigt, wenn eine Sequenz nicht mit einer einzigen Sequenzierung bestimmt werden kann, weil sie zu lang ist.

Nach einer Assemblierung sollte die Qualität der Konsensus-Sequenz überprüft werden (vgl. Abbildung 5-4): Sind Lücken (Insertionen, Deletionen) oder Mismatches (Substitutionen) bei dem Versuch die Sequenzen zusammenzufügen entstanden, sollten die Rohdaten überprüft und wenn nötig korrigiert werden. Dies kann zum Beispiel auftreten, wenn bei der Sequenzierung Kompressionen (vgl. Kapitel 3.2) zu fehlerhaften Ergebnissen beim Base-Calling führen. Wurden Korrekturen durchgeführt, so ist eine erneute Assemblierung notwendig. Ein weiteres Problem, das auftreten kann, ist das Fehlen kompletter Sequenzbereiche. Würde in Abbildung 5-3 die mittlere Sequenz fehlen, könnte man keine Konsensus-Sequenz herstellen. Hier hilft meist nur die erneute Suche in der Datenbank oder eine Sequenzierung um das Problem zu lösen. Die neuen Sequenzen sollten dann erst wieder auf Qualität überprüft werden, bevor sie in die Bibliothek aufgenommen und Assemblierungen durchgeführt werden. Das Variieren der Parameter-Werte kann jedoch ebenfalls zu besseren Ergebnissen führen

und sollte immer zuerst in Betracht gezogen werden. Ab wann eine Konsensus-Sequenz als solche akzeptiert wird, hängt von der jeweiligen Fragestellung ab. Sollen zum Beispiel SNPs detektiert werden, ist es wichtig, jede Lücke oder Substitution genau zu überprüfen. In unserem Projekt, bei dem die Identifizierung von Pilzen im Vordergrund stand, war es vor allem wichtig, dass die Sequenzen sowohl variable als auch konservierte Bereiche enthielten. Das bedeutete, vor allem der ITS-Bereich (für die Generierung von Sonden) und das Ende der 18S - bzw. der Anfang der 28S - Gene (für die Generierung von Primer), mussten vorhanden sein. Es war jedoch nicht notwendig, dass die komplette ribosomale DNA als Konsensus-Sequenz vorlag.
Falls eine Sequenz aus mehreren Komponenten bestand (wie die ribosomale DNA), sollten die Konsensus-Sequenzen zusätzlich noch annotiert, d.h. die Bereiche der einzelnen Komponenten, definiert werden.

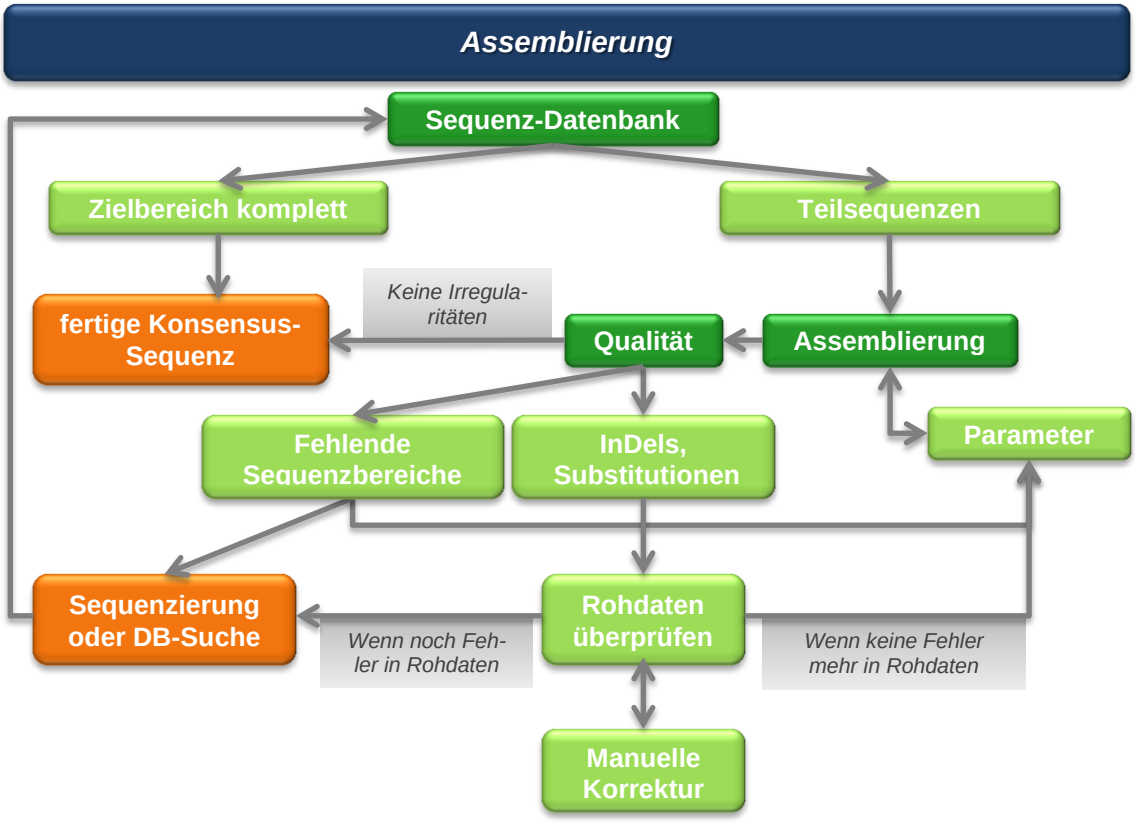

Abbildung 5-4: Zweiter Schritt – Die Assemblierung

Der zweite Schritt beinhaltet die Assemblierung, d.h. das Zusammensetzen von einzelnen Teilsequenzen zu einer Konsensus-Sequenz. Nach der Überprüfung der Qualität und eventuellen Neusequenzierungen bzw. Datenbank-Suchen können die Zielsequenzen abgespeichert und weitere Analysen angeschlossen werden.

5.4 Multiples Sequenz Alignment

In diesem Schritt werden die Sequenzen auf Ähnlichkeiten und Unterscheide untersucht (vgl. Abbildung 5-5). Dazu muss erst geklärt werden, welche Sequenz-Bereiche näher untersucht werden sollen. Ist die komplette, im vorhergehenden Schritt assemblierte Sequenz Ziel der Analyse oder sollen nur kleine Bereiche näher betrachtet werden? Unser Projekt zum Beispiel, benötigt für die Identifizierung von Pilzen Sequenz-Bereiche, die sich von Spezies zu Spezies unterscheiden. Daher würde man nur die variablen Bereiche (z.B. ITS - Sequenzen) in die Untersuchung einbeziehen. Für die Amplifizierung möglichst vieler ITS - Bereiche in verschiedenen Spezies, würde man jedoch eher das Ende der 18S - oder den Anfang der 28S - Sequenz nutzen, da hier konservierte Regionen vorliegen.

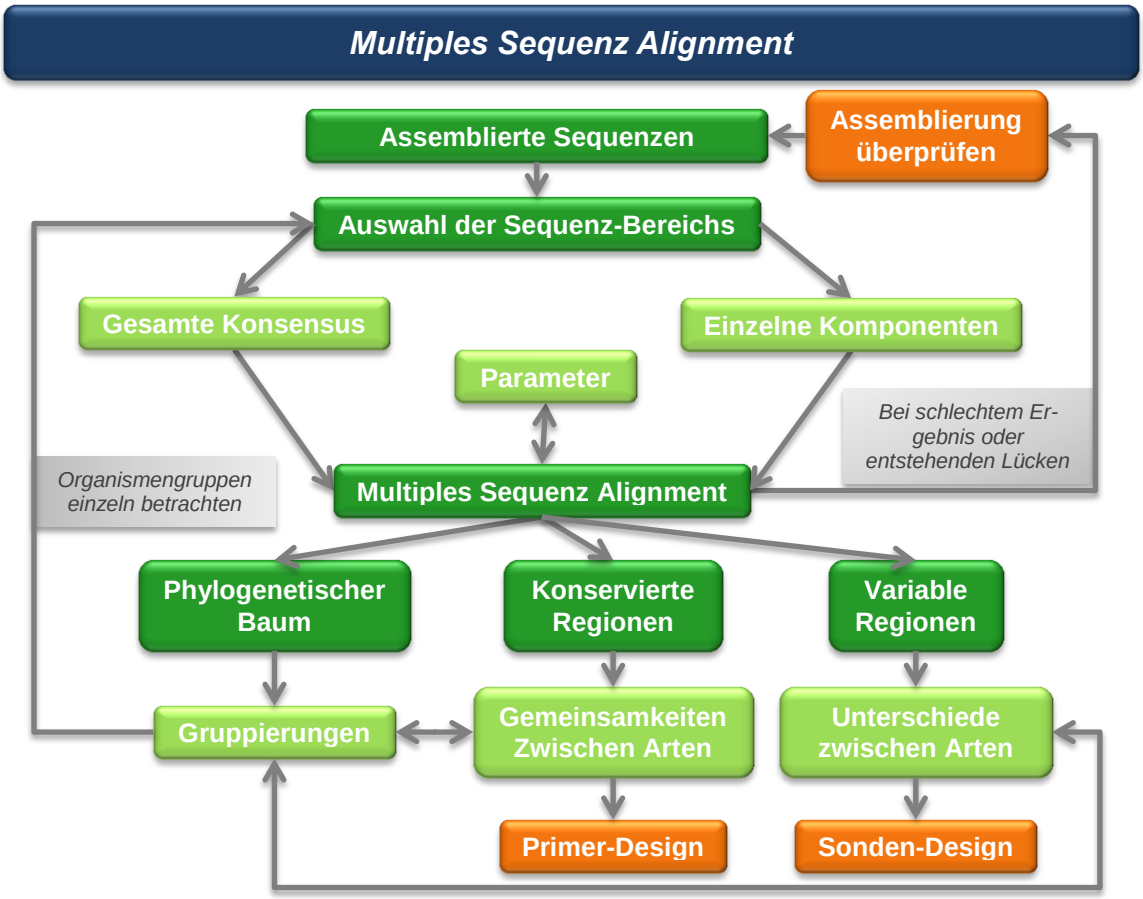

Abbildung 5-5: Dritter Schritt - Multiples Sequenz Alignment

Der dritte Schritt dient der gleichzeitigen Analyse der Sequenzen in der Bibliothek. Ein MSA vergleicht alle Sequenzen miteinander und es sind leicht konservierte und variable Bereiche erkennbar. Des Weiteren geben phylogenetische Bäume Aufschluss über mögliche Gruppierungen von Sequenzen die ähnlicher zueinander sind als zu anderen Sequenzen.

Wurde eine Auswahl getroffen, wurde mit Hilfe eines Programms ein MSA durchgeführt. Auch hier konnten verschiedene Parameter eingestellt und variiert werden um beste Ergebnisse zu erzielen. Ein geeignetes Tool für MSA stellt zum Beispiel ClustalW2, welches auf den Websites des *European Bioinformatics Institutes (EBI)* angeboten wird. Aus einem MSA kann man verschiedenste Aussagen ableiten. Zum einen sind schnell konservierte und variable Bereiche in den Sequenzen erkennbar. Die variablen Bereiche geben Aufschluss über artspezifische Sequenz-Bereiche und aus den konservierten Bereichen lassen sich Gemeinsamkeiten zwischen Arten ablesen. Des Weiteren können phylogenetische Bäume erstellt werden, mit deren Hilfe die Ähnlichkeiten zwischen Sequenzen bildlich dargestellt werden können. Dadurch kann man Organismengruppen, deren Sequenzen besonders ähnlich zueinander sind, einfacher erkennen (vgl. Abbildung 5-6).

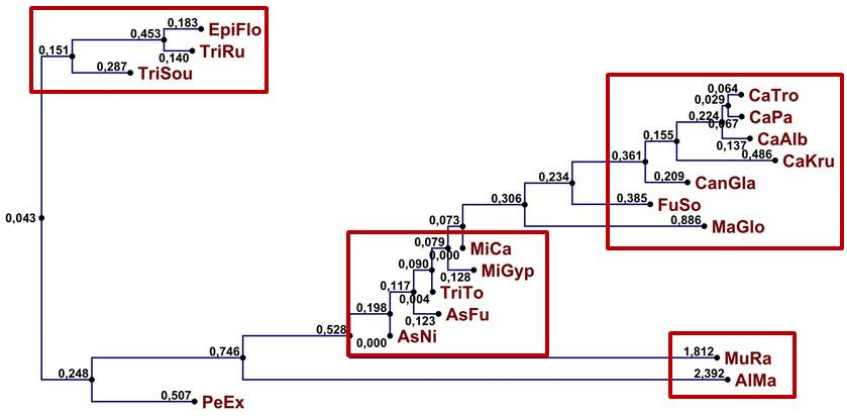

Abbildung 5-6: Phylogenetischer Baum

Mit Hilfe von MSA lassen sich phylogenetische Bäume erstellen, womit die Ähnlichkeiten der Sequenzen zueinander bildlich darstellen werden können. Die rot eingefassten Bereiche stellen daraus abgeleitete Organismengruppen dar. Jeder Bezeichnung (z.B. EpiFlo) repräsentiert eine Seqeunz. Die Zahlen werden als Distanzen bezeichnet und sind ein Maß für die Abstände der einzelnen Sequenzen zueinander.

Zum Beispiel ist es bei der Entwicklung von Primer wichtig, dass die Sequenzen keine zu großen Unterschiede besitzen (d.h. die Gruppen im phylogenetischen Baum dürfen nicht zu stark ausgeprägt sein). Für artspezifische Sonden ist jedoch genau das Gegenteil gefordert. Die Sequenzen sollten hier so viele Unterschiede wie möglich enthalten. Bei einem MSA kann es jedoch auch passieren, dass z.B. sehr große Lücken in Sequenzen eingefügt werden. Dies deutet darauf hin, dass in der Assemblierung Fehlerhafte Konsensus-Sequenzen gebildet wurden und evtl. schon die Rohdaten fehlerhaft sind. Hier sollten die entsprechenden Ausgangsdaten in der Datenbank geprüft werden.

5.5 Primer-/Sonden-Design

Die Entwicklung von Primer hat in der Molekularbiologie einen hohen Stellenwert. Sie werden benötigt um DNA zu vervielfältigen oder zu Sequenzieren. Sonden hingegen werden zur Identifikation von Organismen auf zum Beispiel Microarrays eingesetzt. Da für jede Anwendung spezifische Anforderungen an Primer und Sonden bestehen ist hier nur ein grober Überblick über das Sonden- und Primer-Design gegeben (vgl. Abbildung 5-7).

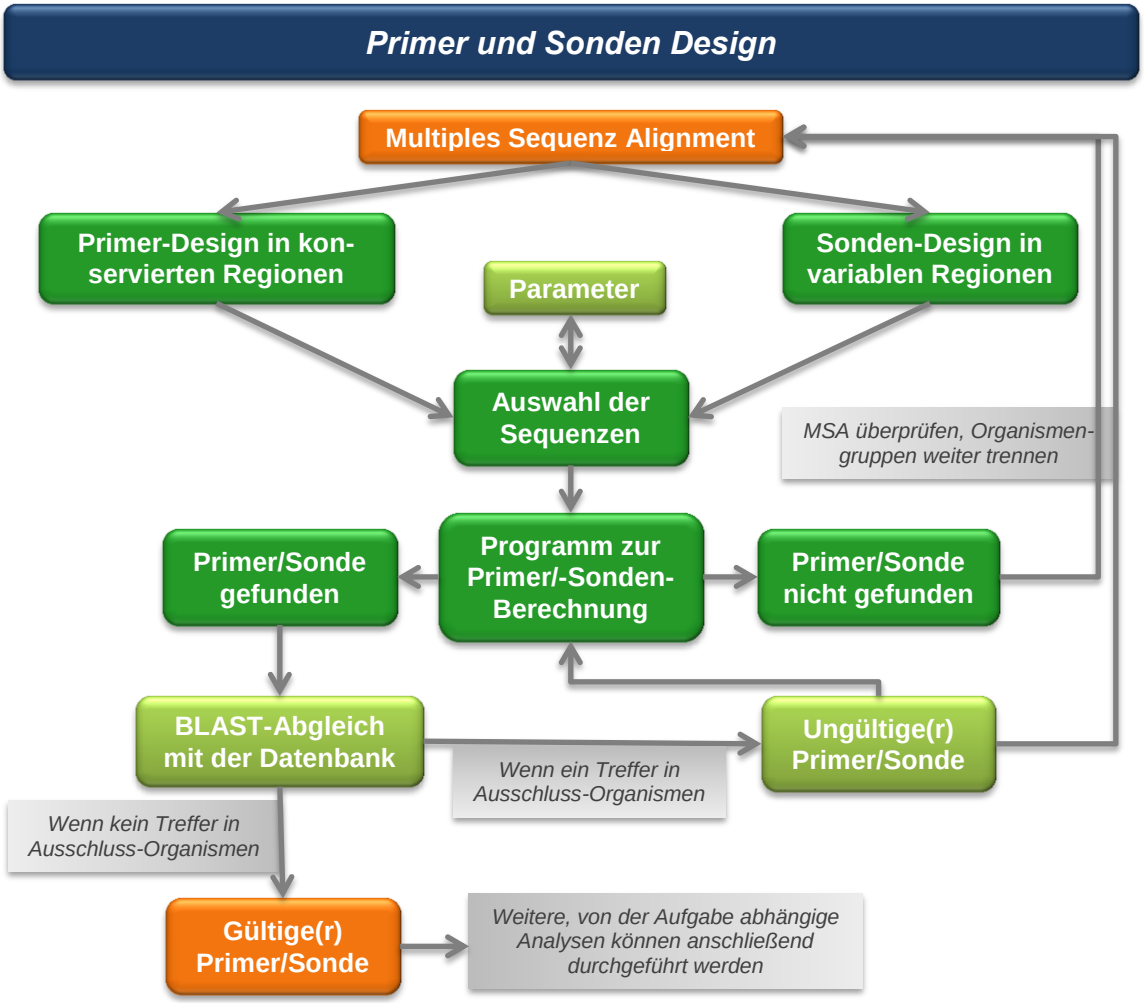

Abbildung 5-7: Vierter Schritt: Sonden- und Primer-Design

Das Primer- und Sonden-Design ist bei vielen molekularbiologischen Anwendungen elementar. Zuerst muss ein Datensatz von Sequenzen erstellt werden, für den Primer oder Sonden generiert werden soll. Ein entsprechendes Programm berechnet mögliche Primer oder Sonden, die anschließend durch einen BLAST-Abgleich mit der Datenbank validiert werden müssen.

Genauere Informationen zu Parametern, Einflussgrößen sowie einen Algorithmus zur Primer-/Sonden-Berechnung werden in der Bachelor-Arbeit von Janine Brettschneider gegeben [Brettschneider, 2011].

Ausgangspunkt ist das MSA, welches im vorherigen Schritt durchgeführt wurde. Man kann entweder Primer für zum Beispiel eine PCR oder Sonden für Detektionen von Organismen herstellen. Je nachdem welches Ziel verfolgt wird, werden entsprechende Sequenzen ausgewählt. Diese werden mit Hilfe eines Programms bearbeitet und anhand der Eigenschaften, die eingestellt wurden, die optimalen Primer bzw. Sonden berechnet. Werden keine Ergebnisse erzielt, so sollte der Datenpool überprüft, neue Sequenzen hinzugefügt oder gelöscht bzw. erneute MSA durchgeführt werden.

Wurden Primer oder Sonden berechnet, sollten sie auf ihre Gültigkeit überprüft werden. Das bedeutet, dass sie nur an die Ziel-Sequenzen, für die sie entwickelt wurden, binden. Um dies zu überprüfen, werden Abgleiche mit großen Datensammlungen vorgenommen. Bei der Identifikation von Pilzen ist es wichtig, dass ein Primer nur Pilze amplifiziert und zum Beispiel nicht an eine Sequenz im menschlichen oder tierischen Genom bindet. Dafür gibt es eine Liste mit Ausschluss-Organismen. Die Validierung wird mit Hilfe von BLAST durchgeführt. Damit werden alle Sequenzen in der Datenbank geprüft, ob die die Primer-/Sonden-Sequenz beinhalten.

Gibt es einen Treffer, so wird der/die Primer/Sonde verworfen und anschließend eine neue Berechnung durchgeführt. Ein Primer/Sonde wird erst als gültig betrachtet, wenn keine Treffer mehr angezeigt werden oder die Treffer nicht in den Ausschluss-Organismen liegen.

6 Qualitätsbeurteilung von DNA-Sequenzen

Für die Entwicklung einer Diagnostik zur Identifizierung von humanpathogenen Pilzen wurden im Rahmen dieses Projektes vor allem DNA-Sequenzen aus den Datenbanken, wie GenBank (NCBI) oder ArbSilva genutzt. Aufgrund des hohen Anteils fehlerhaft annotierter Sequenzen (vgl. Kaptiel 3.4) und der damit verbundenen Fehlerrate bei Datenbanksuchen, soll im folgenden Kapitel eine Möglichkeit vorgestellt werden, selbst sequenzierte Daten, ausgehend von ihren Rohdaten, in ihrer Qualität zu überprüfen. Dafür soll eine Klassifizierung in verschiedene Qualitätsstufen gefunden, vorgestellt und validiert werden. Zum Schluss werden noch bereits vorliegenden assemblierten Datenbank-Sequenzen [Kropp, 2011], die für die Entwicklung eines Diagnostik-Kits genutzt werden sollen, auf Richtigkeit geprüft. Als zweite ergänzende Datenquelle in der nur qualitativ hochwertige DNA-Sequenzen abgespeichert sind, soll BOLD dienen[1] (vgl. Kapitel 4.5).

6.1 Methoden

6.1.1 *Sequenzierung der ITS-Bereiche*

Von insgesamt 100 humanpathogenen Pilzarten (vgl. Anlagen, Tabelle A-1, Spalte „Sequenziert"), welche in der Diagnostik identifiziert werden sollen, wurden von 48 Spezies die ITS-Bereiche sequenziert. Vor der Sequenzierung wurde eine PCR-Amplifikation der ITS-Region (Primer: ITS1 und ITS4, vgl. Tabelle 6-1) und eine anschließender Kontrolle der Amplifikate mittels Agarosegelelektrophorese durchgeführt. Die dabei entstehenden Amplikons waren ca. 550 bp lang. Die Proben wurden danach an ein externes Sequenzier-Labor[2] versandt. Nach einer Aufreinigung, um überschüssige Primer und dNTPs zu entfernen, wurde eine Sequenzierung nach Sanger durchgeführt. Verwendet wurde die *Cycle Sequencing* Technik in Kombination mit der *Big Dye Terminator Chemistry* von *Applied Biosystems* (Foster City, US) sowie eine anschließende Kapillargelelektrophorese unter Benutzung eines *ABIPrism 3730xl Genetic Analyzer* ebenfalls von *Applied Biosystems*. Die Sequenzierung erfolgte bidirektional mit den in Tabelle 6-1 aufgelisteten Primer. Genaue Parameter sind der

[1] Alle verwendeten Daten, für die Auswertung erstellen Excel-Tabellen, Bilder und Literautquellen sind auf der beigefügten CD zu finden.
[2] LGC Genomics in Berlin (https://shop.lgcgenomics.com/index.php)

Veröffentlichung von Heiner zu entnehmen. [Heiner, 1998] Die erhaltenen Rohdaten wurden mit Hilfe der *Sequencing Analyzer Software for ABI Prism® BigDye® v3 Chemistries* auf Windows NT Basis ausgewertet. Die Rohdaten sind im *.ab1-Format* auf der beigelegten CD abgespeichert.

Tabelle 6-1: Primer für die PCR und Sequenzierung

Die Tabelle zeigt die für die PCR und das *Cycle Sequencing* benutzten Primer sowie deren Sequenz, Länge, und Annealingtemperatur (T_A). „Forward" steht für die Sequenzierung des Sense und Reverse für die Sequenizerung des Antisense-Stranges der DNA. Die Abkürzungen stehen für die Organismen: *Scopulariopsis brevicaulis (ScoBre)*, *Engyodontium album (EngAlb)* und *Trichoderma viride (TriVir)*

Primer Name	Primer Sequenz	Länge in bp	T_A in °C	Bemerkung
ITS1	TCCGTAGGTGAACCTGCGG	19	55	Forward
ITS4	TCCTCCGCTTATTGATATGC	20	50	Reverse
ITS1-2W	TCCGTWGGTGAACCWGCGG	19	51	Forward, für: *ScoBre*; *Chaetomium spp.; EngAlb; TriVir*

6.1.2 *Parameter für die Analyse der DNA-Sequenzen*

Die Analyse der sequenzierten ITS-Bereiche wurde mit der Software *CodonCodeAligner V. 3.7.* durchgeführt. Es wurde die frei zugängliche, 30-tägige Testversion genutzt. Das Programm beinhaltet eine Implementierung des PHRED-Algorithmus wodurch die Qualität der einzelnen Nucleotide akkurat berechnet werden kann (vgl. Kapitel 3.3.2). Diese Funktion ist im Menü unter *Samples* > *Call Bases* implementiert. Um qualitativ schlechte Basen (PHRED < 20) am Anfang und Ende einer Sequenz abzutrennen wird anschließend ein EndClipping[1] (*Samples* > *EndClipping*) durchgeführt. Für Anfang und Ende der Sequenz wurden unterschiedliche Einstellungen gewählt, da der Anfang weniger qualitativ schlechte Basen besitzt als das Ende eines Trace, wo die Signalstärken kontinuierlich schwächer werden (vgl. Kapitel 3.1.3).

- Anfang: Fehlerrate kleiner 0,01 (entspricht PHRED 20) in einem Fenster von 20 Basen
- Ende: Fehlerrate kleiner 0,01 in einem Fenster von 50 Basen.

[1] Entfernen von qualitativ schlechten Abschnitten am Anfang und Ende einer Sequenz

Anschließend wurde eine Assemblierung (im Menü unter *Contig > Assembly*) der jeweiligen Forward- und Reverse- Sequenzen durchgeführt. Der verwendete Algorithmus des Programms *CodonCodeAligners* ist fast identisch mit dem PHRAP-Algorithmus[1], der auf Grundlage der berechneten PHRED-Werte identische oder ähnliche Bereiche in zwei oder mehr Sequenzen sucht um diese zu alinieren. Die enstandenen Konsensus-Sequenzen wurden hinsichtlich Lücken (Insertionen, Deletionen), Mismatches(vgl. Kapitel 3.2.6) und Anzahl der Ns untersucht. Die Anzahl der Ns stammt aus den Originalsequenzen und stellt eine Sonderform des Mismatches dar.

6.1.3 *Parameter für die Klassifizierung*

Ein weiterer Schritt der Sequenzanalyse beinhaltet die Klassifizierung der Sequenzen nach ihrer Qualität. Die Anforderungen die an die Klassifizierung gestellt werden sind folgende:

- keine Sequenz darf in ihrer Qualität höher bewertet werden, als sie ist
- der Anwender muss, auch wenn er keine Kenntnis über die Parameter hat, schnell erkennen können ob eine Sequenz für weitere Analysen gut genug ist oder nicht (z.B. das Ampelsystem: rot bedeutet sehr schlecht, grün bedeutet keine Fehler oder sehr gut und bei gelb könnten Fehler vorhanden sein).
- es muss eine gewisse Sensitivität gegeben sein, d.h. das zu bewertende Objekt darf nicht aufgrund kleinerer Fehler als sehr schlecht eingestuft werden.

Um geeignete Parameter, die diese Punkte erfüllen, zu finden, dienen zum einen die anerkannten BOLD-Kriterien (Tabelle 4-1) und zum anderen der Barcode-Datenstandard (vgl. Kapitel 4.4 und 4.5) [Hanner, 2009], [Ratnasingham, 2007] als Grundlage.

Daraus ergeben sich vier Kategorien: *high, medium, low* und *failed*. Eine Sequenz wird als hochqualitativ – *„high“* - bezeichnet, wenn der PHRED-Mittelwert größer 40 ist, weniger als 1 % Ns vorhanden sind. Eine als *„medium“* eingestufte Sequenz besitzt einen Mittelwert von 30 – 40 sowie 1 % bis 2 % Ns und eine Sequenz mit der Bewertung „*low*“ den Mittelwert von 20 – 30 und 2 % bis 4 % Ns. Eine Sequenzierung wird als fehlgeschlagen – „*failed*“ – bezeichnet, wenn keine Sequenz aus dem Trace abgeleitet werden kann (PHRED-Mittelwert < 29) oder mehr als 4 % Ns vorhanden sind. Als „N“ werden alle Basen PHRED < 20 bezeichnet. Kann eine Sequenz anhand dieser Kriterien nicht eindeutig bewertet werden, zum Beispiel wenn der Durchschnitt

[1] nähere Informationen unter http://www.phrap.org/phredphrapconsed.html#block_phrap

PHRED > 40, die Anzahl der Ns aber zwischen 2 % und 4 % liegt, wird der Anteil qualitativ hoher Basen (PHRED > 30) in einer Sequenz als weiteres Kriterium überprüft. Dabei wird ein Schwellwert von 75 % genutzt. In Tabelle 6-2 sind die Einteilungen dargestellt.

Tabelle 6-2: Zusatzkriterien für Klassifizierung der Sequenzen

Dargestellt sind die, für die Klassifizierung notwendigen Zusatzkriterien, um Sequenzen in eine der vier Kategorien *high, medium, low* oder *failed* einzuteilen. Dabei wird nicht nur auf den Mittelwert der PHRED-Werte und die Prozentzahl qualitativ schlechter Basen (Ns entsprechen Basen mit PHRED < 20) geachtet, sondern zusätzlich auf das Verhältnis qualitativ sehr guter Basen (PHRED > 30) zu Sequenzlänge.

Einstufung PHRED - Mittelwert	Einstufung Anteil Ns	Anteil PHRED > 30 Basen > 75 %	Anteil PHRED > 30 Basen < 75 %
> 40 (h)	1 % - 2 % (m)	high	Medium
> 40 (h)	2 % - 4 % (l)	medium	Low
> 40 (h)	> 4 % (f)	medium	Low
30 -40 (m)	2 % - 4 % (l)	medium	Low
30 -40 (m)	> 4 % (f)	low	Failed
20 – 30 (l)	> 4 % (f)	low	failed

6.1.4 *Filtern der Barcodes aus BOLD*

Alle Pilz-Sequenzen, welche von BOLD heruntergeladen wurden, stammen aus den Rubriken *Fungal Barcoding* oder *GenBank Fungi.* Die Daten wurden nach Projekten und Sequenzart (ITS, COI, 18S oder 28S) gegliedert abgespeichert. Zusätzlich wurden die zugehörigen Spreadsheets sowie Trace-Files lokal gespeichert. Für den Download der Sequenzen und des Spreadsheets wurden die in Abbildung 6-1 dargestellten Einstellungen vorgenommen. Die Trace-Files wurden nach ihren Kategorien (*high, medium, low* und *failed*) gegliedert abgelegt.

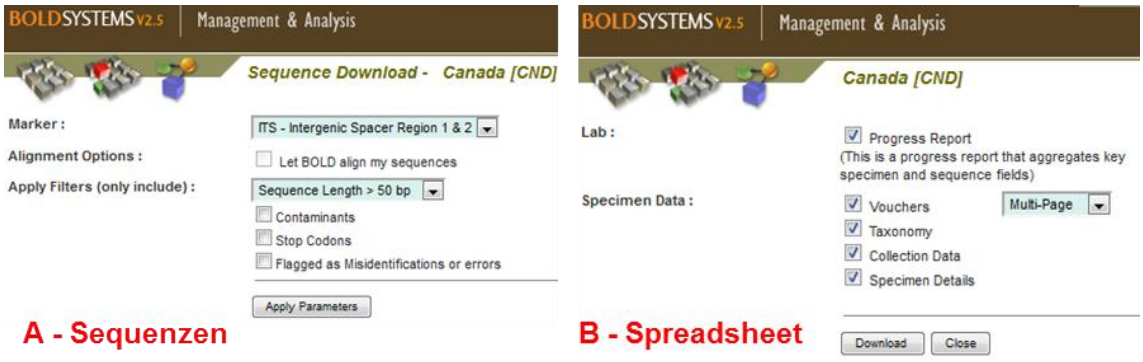

Abbildung 6-1: DownloadEinstellungen in BOLD

Der Download der Sequenzen (A) wurde nach den abgebildeten Standrad-Einstellungen vorgenommen. Bei einigen Projekten konnten im Feld „Marker“ auch die Sequenzen für COI, 18S und 28S ausgewählt werden. Diese wurden ebenfalls gespeichert. Die Einstellungen für den Download der Spreadsheets sind in B dargestellt.

6.2 Analyse der Elektropherogramme

Um mögliche Fehler beim *Base-Calling* oder dem Assemblieren besser bewerten zu können, wurden zu Beginn alle Elektropherogramme manuell analysiert und auf mögliche Fehlerquellen, die in Kapitel 3.2 erläutert wurden, untersucht. Insgesamt standen für jede Pilz-Art zwei Traces zur Verfügung. Ausnahmen hierbei waren *AcrMur, ChaMur* und *PenPic*, für die jeweils vier Traces (zweimal Forward und Reverse) vorlagen. Daraus ergaben sich insgesamt 102 zu untersuchende Elektropherogramme.
In insgesamt 50 der 102 Traces konnten Irregularitäten[1] entdeckt werden. Zuerst fiel auf, dass in allen EPG das Signal bei durchschnittlich 550 nt abbricht und keine Peaks mehr erkennbar sind. Dies war jedoch mit der durchgeführten PCR vor der Sequenzierung zu erklären, wodurch die DNA-Sequenzen eine vordefinierte Länge besaßen und in einem Trace nur bis zu einer bestimmten Länge Peaks auftreten können. Diese Beobachtung wurde daher nicht als „Fehler“ betrachtet. Die am häufigsten auftretenden Irregularitäten waren *Dye Blobs* (in 41 der 50 EPG). Diese haben unterschiedliche Ausprägungen. Es können *Cytosin (C-) Blobs*, *Thymin (T-) Blobs* oder auch gemischte *Blobs* mit mehr als einer Base auftreten (vgl. Abbildung 6-2). Die Höhen der Peaks können ebenfalls variieren. Es gab sehr deutliche *Dye Blobs*, deren Höhe mehr als das doppelte der normalen Peakhöhe beträgt (vgl. Abbildung 6-2 links). Im Gegensatz dazu konnten sie jedoch auch sehr niedrig sein und als sehr breite, kleine Peaks auftreten (vgl. Abbildung 6-2 rechts). Die Höhe eines *Blobs* hat vor allem einen Einfluss auf die Genauigkeit und Fehlerfreiheit des *Base-Callings*. Während große *Blobs* eher von der Software als solche erkannt, könnten kleine als mögliche sekundäre Peaks zum Beispiel zu einer falsch vorhergesagten Base führen [Cheng, 2008].

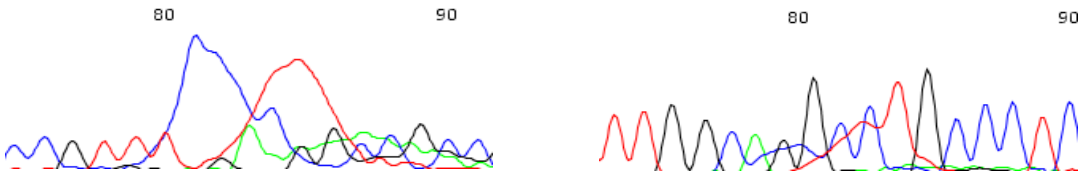

Abbildung 6-2: Stark und schwach ausgeprägte Dye Blobs

In einigen Traces treten Dye Blobs im Bereich von 80 bis 120 nt auf. Diese können unterschiedlich stark ausgeprägt sein. Abgebildet ist ein doppelter Dye Blob der Basen C (blau) und T (rot) die deutlich über der Signalstärke der restlichen Peaks liegen (links) und ein relativ schwacher Dye Blob, welcher jedoch auf das Base-Calling negative Einfluss nehmen kann (rechts).

[1] Der Begriff Irregularität wird in diesem Kapitel für Beobachtungen in EPG verwendet, die nicht mit einem idealen Trace übereinstimmen und so zu Fehlern im Base-Calling oder anderen Sequenzanalytischen Schritten führen können.

Als zweithäufigste Irregularität trat Signalschwäche auf. Die jeweiligen Signalintensitäten wurden mit Hilfe von *CodonCodeAligner* ausgelesen[1]. Die Bezeichnung „stark signalschwacher Trace“ wird verwendet wenn mindestens zwei Basen eine Signalstärke kleiner 100 besitzen und „signalschwacher Trace“ (17 EPG) wenn mindestens zwei Basen Signalstärken zwischen 100 und 150 (7 EPG) aufweisen [Nucleics, 2010], [Dasenko, 2011]. Der Trace von *FusSol R* bildete hier eine Ausnahme. Er besaß am Anfang des Trace relativ gute Signalstärken und ab einer Base von 150 fallen diese plötzlich auf ca. 1/5 der vorherigen Stärke ab. Die Folge könnte eine ungenaue, qualitativ schlechte oder nicht ablesbare DNA-Sequenz sein, da durch die geringere Signalstärke das Signal-Rausch-Verhältnis zunimmt und dadurch die Peaks schlechter den entsprechenden Basen zugeordnet werden können. Eine geringe Signalstärke war auch im Elektropherogramm erkennbar, wenn man einen Trace mit guter und schlechter Signalstärke direkt miteinander verglich (vgl. Abbildung 6-3).

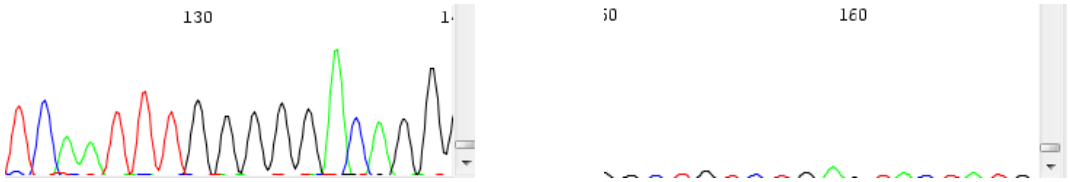

Abbildung 6-3: Vergleich zwischen signalstarkem und signalschwachem Trace

Nach dem die Traces gleich skaliert wurden (vertikale ScrollBar auf die niedrigste Stellung gesetzt), konnte die unterschiedliche Stärke der Signale gewertet werden. Ein signalstarker Trace (links) zeigt klar erkennbare Peaks wobei bei einem signalschwachen nur schwer deutliche Peaks erkennbar sind (rechts).

Weiterhin wurden Traces identifiziert, bei denen die Peaks um N-1 (nach links) oder um N+1 (nach rechts) verschoben wurden und einen deutlichen sekundären Peak erzeugten. Der Shift der Peaks um N-1 deutet auf eine fehlerhafte Primer-Synthese hin, bei der ein Teil der Primer für die Sequenzierung um einen Nukleotid kürzer sind. Dadurch unterscheiden sich auch die Fragmente in ihrer Länge um einen Nukleotid und es treten Doppelpeaks auf (Abbildung 6-4). Bei der Sequenz von *AcrMur F2* kann am Anfang des Trace eine Verschiebung der Peaks um N+1 beobachtet werden. Je weiter rechts man den Trace betrachtet, umso weniger stark ist dieses Phänomen ausgeprägt. Man könnte dies ebenfalls als Primerfehler betrachten. Doch da die Ausprägung der Sekundarpeaks kontinuierlich abnahm und damit nicht dem typischen Bild eines N-1 Primers entsprach (vgl. Abbildung 6-4), wurde dieser Trace bei der Irregularität „Kontamination“ eingeordnet.

[1] Die Signalstärken eines Trace werden mit Hilfe der Funktion „Samples > Sample Information“ ausgelesen. Sie sind nach Art der Base getrennt aufgelistet z.B. „SIGN=A=297,C=342, G=300, C=374“.

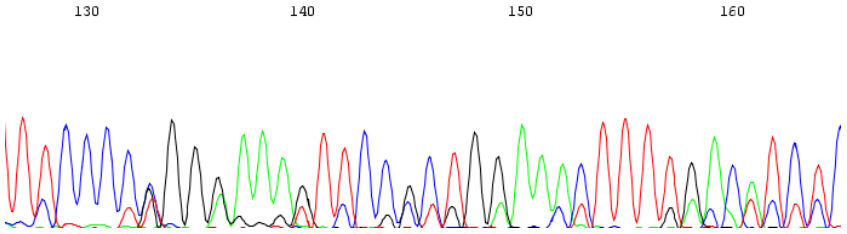

Abbildung 6-4: Trace mit Primerfehler

Bei der Synthese der verwendeten Primer für eine Sequenzierung ist es möglich, dass ein Nukleotid zu wenig oder zu viel eingebaut wird. Dadurch ist der fehlerhafte Primer um ein Nukleotid kürzer und verursacht im Trace um einen Nukleotid verschobene Peaks.

In zwei Traces konnte eine wellenartige Anordnung der Peaks nach einem längeren Mononukleotid-Abschnitt beobachtet werden. Dies könnte durch einen *Slippage* der DNA-Polymerase verursacht worden sein.
Eine mögliche Bildung von Chimären konnte bei den EPGs von *SynRac* beobachtet werden. Nach den ersten 100 (Reverse) bzw. 400 (Forward) Peaks, die eindeutig erkennbar sind, vermischen sich die Signale stark, was durch eine Rekombination der Target-Sequenz hervorgerufen werden kann (Abbildung 6-5).

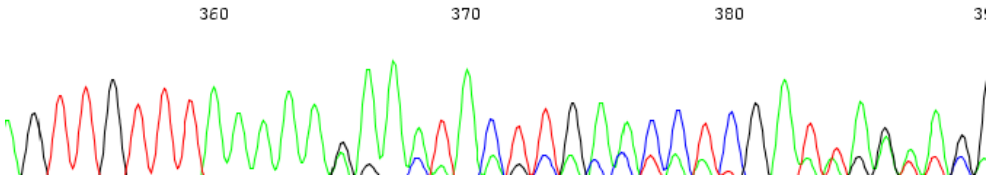

Abbildung 6-5: Auftreten von Chimären im Trace

Die zwei Traces von *SynRac* besitzen am Anfang einen klaren, eindeutigen Trace. Ab einem bestimmten Peak treten jedoch vermischte Signale auf. Dies äußert sich vor allem Doppelpeaks oder Sekundärpeaks. Als mögliche Ursache könnten DNA-Chimären, die durch Rekombination entstanden sind, genannt werden.

Weiterhin wurden bei der Forward- und Reverse- Sequenz von *EngAlb* relativ hohe und deutliche Sekundärpeaks im kompletten Trace beobachtet. Dies könnte auf eine Verunreinigung mit Fremd-DNA hindeuten.
Ein weiteres Indiz dafür war, dass nach Base 530 bei *EngAlb R* bzw. 534 bei *EngAlb F,* wo der Trace und damit die Sequenz normalerweise enden sollte, noch relativ deutliche, wenn auch schwache Signale auftraten (Abbildung 6-6). Dies würde bedeuten, dass die kontaminierende DNA länger ist als das eigentliche Target und die Sequenzierung längere Fragmente erzeugen würde. Die Folge wäre eine falsch geschlussfolgerte DNA-Sequenz nach dem Base-Calling.

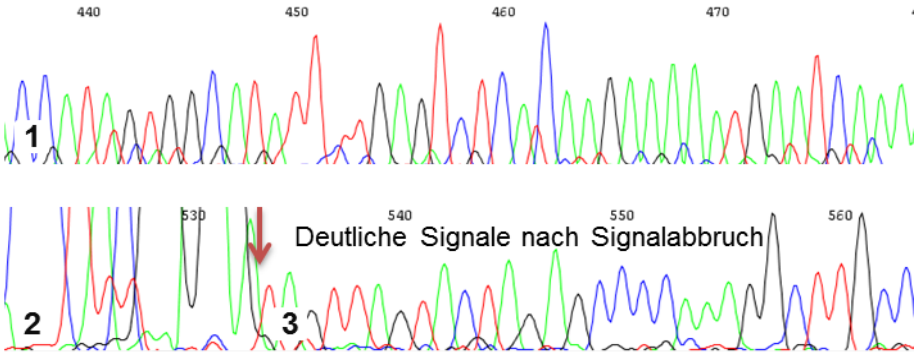

Abbildung 6-6: Sekundärpeaks in einem Trace

Der abgebildete Trace von *EngAlb R* besitzt teilweise hohe Sekundärpeaks (1), die auf eine Verunreinigung hindeuten könnten. Diese Vermutung wird durch das Auftreten klarer (jedoch schwacher) Signale (3) nach dem regulär beobachteten Signalabbruch bei ca. 330 nt (2, roter Pfeil) bekräftigt

Die Verteilung aller Irregularitäten ist in Abbildung 6-7 noch einmal zusammengefasst und in Tabelle A-4 im Anhang detailliert dargestellt.

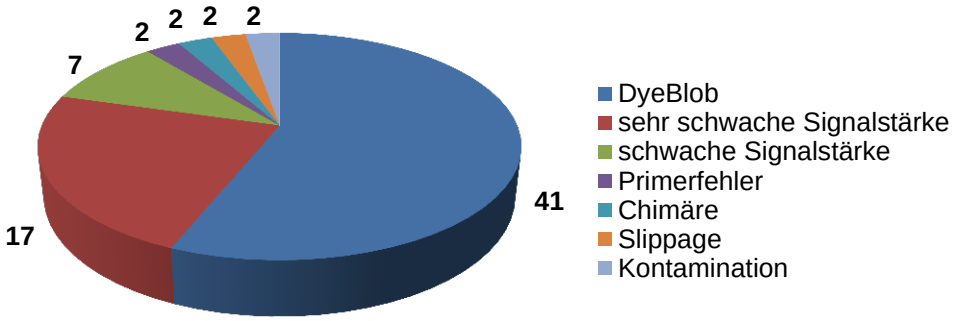

Abbildung 6-7: Verteilung der Irregularitäten

Das Diagramm stellt zusammenfassend die Anzahl der beobachteten Irregularitäten in den untersuchten 102 Traces dar. Insgesamt wurden in 50 Traces Irregularitäten entdeckt, wobei in einigen Traces mehrere Irregularitäten beobachtet wurden (z.B. Signalschwäche und *Dye Blob*).

6.3 Analyse der DNA-Sequenzen

Die *Base-Calling* Funktion des Programms *CodonCodeAligner* konvertiert die EPG in DNA-Sequenzen. Gleichzeitig wird jeder Base (A, T, G oder C) ein Qualitätswert (PHRED-Wert) zugeordnet. Je höher dieser Wert ist, desto wahrscheinlicher ist es, dass die Base richtig vorhergesagt (vgl. Kapitel 3.3) worden ist. Diese Werte werden vom Programm in Text-Dateien, mit der Endung „phd.1“ gespeichert[1] und können mit Hilfe eines einfachen Text-Editors geöffnet werden. Zu jedem EPG existiert eine solche Datei. Um die Daten zur Analyse besser in eine Excel-Tabelle importieren zu können, wurde ein kleines JAVA-Programm (Quellcode im Anhang) entwickelt. Es liest alle phd.1-Dateien in einem Ordner aus und speichert die benötigten Daten (Basen-Abfolge mit zugehörigen Qualitätswerten und Name der Datei) in eine Text-Datei ab.
Nach dem *Base-Calling* ergaben sich mittlere Sequenzlängen von 1385 bp. Jedoch ist der Anteil der qualitativ schlechten Basen (PHRED <20) mit durchschnittlich 765 bp relativ hoch. Aus diesem Grund wurde die End-Clipping Funktion genutzt, um nur den mittleren, qualitativ hochwertigen Bereich einer Sequenz zu isolieren (vgl. Kapitel 3.2.4).

6.3.1 *Einfluss von Irregularitäten auf die Sequenzlänge*

Die resultierenden Sequenzen unterschieden sich nun zum Teil stark in ihrer Länge. Es gab insgesamt 13 Sequenzen, die kürzer als 10 bp sind, 7 Sequenzen mit Längen zwischen 200 und 450 bp, 78 Sequenzen zwischen 450 und 600 bp und 4 Sequenzen mit mehr als 600 bp. Bei dem Vergleich der Ergebnisse mit den Analysen der EPG, fällt auf, dass die zugehörigen Traces der 13 kurzen Sequenzen als stark signalschwache eingestuft wurden (*ChaMur F2+R1+R2, FusCul F, OidGri F+R, PenPic F1+F2+R1+R2, ScoFus F+R, TriHar R*). Bei den 7 Sequenzen zwischen 300 und 450 bp handelt es sich um *AspFum F+R, ChaMur F1, FusSol R, GeoCan F+R und RhiSto R*. Bis auf *GeoCan F+R*, die eindeutige EPG und sehr gute Signalstärken besaß, traten bei den anderen fünf Traces Irregularitäten auf, wodurch die kurze Sequenzlänge erklärt werden kann:

- *RhiSto R:* ab Base 150 tritt ein Slippage auf, wodurch keine Sequenz mehr akkurat bestimmt werden kann.

[1] Die Abspeicherung erfolgt in dem Verzeichnis, wo das Programm installiert wurde unter dem Pfad …\CodonCodeAligner\Projects\“Name des Projektes“\phd_dir.

- *AspFum F+R:* Sie wiesen stetig sinkende Signalstärken auf, wodurch das zunehmende Rauschen beim Base-Calling zu qualitativ schlechten und somit zu kurzen Sequenzen führte ebenso wie bei *ChaMur R1*. Die Traces dieser Sequenzen wurden in Tabelle A-4 ebenfalls als (stark) signalschwach eingestuft.
- *FusSol R:* Ab Base 150 nahm die Signalstärke plötzlich ab, was zu schlechteren PHRED-Werten führte.

Die 78 Sequenzen mit Sequenzlängen zwischen 450 und 600 nt besitzen eine mittlere Länge von 516 nt und erreichen damit die ungefähre Größe des ITS-Bereichs. Die Sequenzen von *AbsCor F+R* und *EngAlb F+R* wurden wegen ihrer relativ großen Abweichung von 150 nt zum Mittelwert genauer betrachtet: Die Traces von AbsCor wiesen keine Unregelmäßigkeiten auf und besaßen sehr gute Signalstärken. Daher kann ein Fehler beim Base-Calling und EndClipping ausgeschlossen werden. Bestätigt wird die Sequenzlänge von 790 nt durch die Studie von Garcia-Hermoso und Kollegen, die die Länge des ITS –Bereichs von *AbsCor* zwischen 763–770 nt definieren [Garcia-Hermosa, 2009]. Bei *EngAlb ist* ebenfalls eine sehr gute Signalstärke beobachtet worden. Jedoch ist wahrscheinlich eine, bei der Sequenzierung aufgetretene, Verunreinigung Ursache für Sekundärpeaks die auch über das eigentliche Ende der Sequenz von *EngAlb* detektiert und in Basen konvertiert werden konnten. Dadurch entstand eine längere, jedoch mit Fremd-DNA kontaminierte Sequenz (vgl. Abbildung 6-6). Um diese abzutrennen, wurden die Sequenzen von *EngAlb F+R* nur bis Base 499 bzw. 504 betrachtet. In Tabelle 6-3 sind noch einmal die Ergebnisse der Analyse der Sequenzlängen zusammengefasst.

Tabelle 6-3: Verteilung der Sequenzen hinsichtlich ihrer Sequenzlänge

Nach dem EndClipping kann eine Unterscheidung der Sequenzen hinsichtlich ihrer Sequenzlänge unternommen werden. Es gibt vier Kategorien: sehr kurze (<10), kurze (300-450), der Länge des ITS-Bereichs entsprechende (450-600) und sehr lange Sequenzen (>600). Die Bemerkungen geben zusätzlich wichtige Hinweise zu der jeweiligen Einteilung.

Länge	Sequenzen	Bemerkung
< 10	13	Besitzen alle stark signalschwache Traces
300-450	6	bis auf GeoCan F+R ist kurze Sequenz durch Irregularität entstanden
450-600	80	Anzahl nach dem manuellen Trimmen von EngAlb F+R aufgrund von Kontamination (davor 78)
>600	2	Anzahl nach dem manuellen Trimmen von EngAlb F+R aufgrund von Kontamination (davor 4)

6.3.2 *Klassifizierung der Sequenzen*

Die Qualitäten von Sequenzen können sich, so wie ihre Längen, ebenfalls stark unterscheiden. Mit Hilfe der in Kapitel 6.1.3 beschriebenen Verfahren zur Klassifizierung von Sequenzen wurden folgende Ergebnisse (vgl. Tabelle 6-4) erzielt: Insgesamt konnten 55 Sequenzen mit Status *„high“,* 25 mit *„medium“,* 2 mit *„low“* und 20 mit *„failed“* klassifiziert werden. Davon sind, ohne die in Tabelle 6-2 beschriebenen Zusatzkriterien, 12 Sequenzen als *„high“* und 13 Sequenzen als *„failed“* eingeteilt worden.

Tabelle 6-4: Ergebnis der Klassifizierung der Sequenzen

Es sind die Ergebnisse der Klassifizierung der Sequenzen nach den genannten Kriterien in Kapitel 6.1.3 dargestellt. Es wird gezeigt, wie viele der Sequenzen jeweils mit und ohne Zusatzkriterien zugeordnet werden konnten. Als Zusatzkriterien werden die in Tabelle 6-2 aufgelisteten Daten bezeichnet.

	Anzahl ohne Zusatzkriterien	Anzahl mit Zusatzkriterien	Anzahl gesamt
high	12	43	55
medium	0	25	25
low	0	2	2
failed	13	7	20

Bei einem Vergleich der Einstufungen mit den bisher gemachten Beobachtungen in Kapitel 6.2 und 6.3 fällt auf, dass alle als „*high“* eingestuften Sequenzen keine Irregularitäten im zugehörigem Trace besitzen und bis auf *GeoCan F* (Länge von 309 bp) eine Sequenzlänge größer 450 nt besitzen.

Dagegen konnte bei jeder *„failed“*-Sequenz eine Irregularität entdeckt werden, die sich auf einen Großteil der Sequenz auswirkt. Bis auf *PenGri R* gehören darunter alle Sequenzen mit stark signalschwachen Traces (insgesamt 16, vgl. Tabelle A-4) sowie *SynRac F*, bei der ab Base 130 eine Chimäre Sequenz vermutet wird und *RhiSto F* sowie *FusSol R*, die beide einen signalschwachen Trace besitzen. Zusätzlich wurde bei *RhiSto F* ein Polymerase-Slippage und bei *FusSol R* eine starke Signalabschwächung ab Base 150 beobachtet.

Die zwei als *„low“* klassifizierten Sequenzen sind *RhiSto R*, dessen Trace ebenfalls ein Polymerase-Slippage zeigt und *EngAlb R*, wo relativ hohe Sekundärpeaks ab der Base 350 auftreten. Betrachtet man die „*medium“*-Sequenzen im Gesamten, kann man feststellen, dass hier Sequenzen sowohl mit als auch ohne Irregularitäten eingestuft worden sind. Jedoch ist die Intensität der jeweiligen Irregularität, insofern eine vorhanden ist, bei *„medium“*-Sequenzen schwächer ausgeprägt als bei *„low“*- oder *„failed“*-

Sequenzen. Zum Beispiel trat bei *SynRac R* (*„medium“)* erst ab Base 430 eine Chimäre auf, wodurch im Gegensatz zu *SynRac F* (*„failed“*, Chimäre ab Base 130) eine höhere Chance auf qualitativ höher bewertete Basen bestand. Des Weiteren war bei *EngAlb F („medium“)* die Ausprägung der Sekundärpeaks weitaus schwächer (obwohl die Signalstärke geringer ist) als bei *EngAlb R („low“)*, wodurch die Basen ebenfalls einen höheren PHRED-Wert besaßen (vgl. Abbildung 6-8).

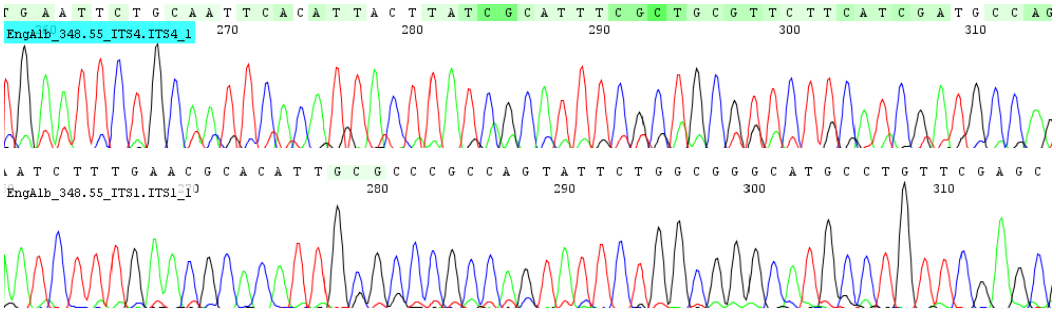

Abbildung 6-8: Unterschiedlich hohe Sekundärpeaks bei *EngAlb F* und *EngAlb R*

Eine Irregularität bei einer, als *„medium“* eingestuften, Sequenz ist in ihrer Intensität schwacher als eine Sequenz mit dem Status *„failed“* oder *„low“.* In der Abbildung sind beispielhaft die Traces von *EngAlb R*, eingestuft als *„failed“,* (oben) und *EngAlb F*, bewertet als *„medium“* (unten) dargestellt. Man erkennt, dass die Höhe der Sekundärpeaks im oberen Trace stärker ausgeprägt ist und die Sequenz schlechtere PHRED-Werte besitzt als die untere (dunkles grün bedeutet PHRED <20, je heller desto höher ist der PHRED-Wert).

Zusammenfassend betrachtet, kann man zwischen den beobachteten Irregularitäten und den jeweiligen Klassifizierungen der daraus geschlussfolgerten Sequenzen einen Zusammenhang erkennen. Je stärker eine Irregularität ausgeprägt war, desto schlechter wurde eine Sequenz eingestuft. Im Anhang unter Tabelle A-3 können die Klassifizierungen und der dafür notwendiger Parameter jeder Sequenz einzeln eingesehen werden.

6.4 Analyse der Assemblierungen

Um fehlerhaft vorhergesagte Basen beim Base-Calling überprüfen und berichtigen zu können, wurde eine Assemblierung der Forward- und Reverse- Sequenzen durchgeführt. Es entstanden 41 Konsensus-Sequenzen (auch *„Contigs*“ genannt). Bei *ChaMur, FusCul, OidGri, PenPic, RhiSto, ScoFus und TriHar* konnten keine Konsensus-Sequenzen erzeugt werden. Der Grund dafür ist, dass bei *RhiSto* der DNA-Slippage zu einer komplett falschen Sequenz geführt hat und bei allen anderen fehlgeschlagenen Assemblierungen mindestens eine Sequenz (Forward oder Reverse oder beide) eine Länge kleiner zehn besitzt und als *„failed“* eingestuft worden ist. Dadurch lassen sich keine überlappenden Bereiche mehr finden. Gegenteilig dazu besitzen die Konsensus - Sequenzen von *AbsCor, AltAlt, BotCin, FusOxy, MucPlu, MucRac, PenMar* und *PenVer* weder Lücken, noch Mismatches noch Ns. Dabei fällt auf, dass alle Forward - und Reverse - Sequenzen in die Kategorie *„high“* eingestuft worden sind. Alle weiteren Assemblierungen, bei denen Fehler in Form von Lücken, Mismatches oder Ns beobachtet werden konnten, wurden manuell überprüft und, wenn möglich, korrigiert. Die Verteilung der Fehler vor und nach dieser Korrektur ist in Abbildung 6-9 und eine detaillierte Aufstellung der Fehler in Bezug auf die jeweiligen Konsensen in Tabelle A-5 und Tabelle A-6 im Anhang dargestellt.

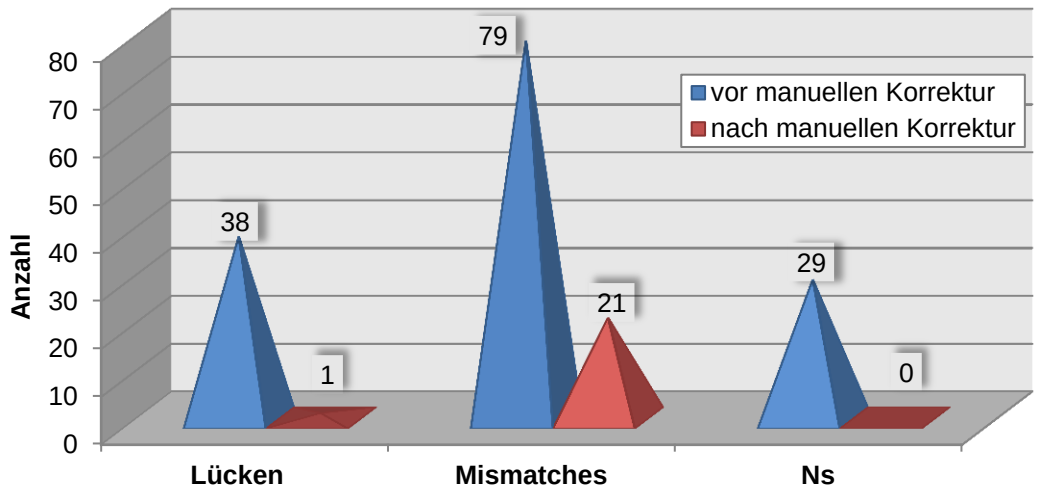

Abbildung 6-9: Anzahl der Lücken, Mismatches und Ns vor und nach der manuellen Korrektur

Das Diagramm gibt die Anzahl der Fehler aufgeteilt in Lücken, Mismatches und Ns in den Assemblierungen wieder. Dabei wird die Anzahl vor und nach der manuellen Korrektur dargestellt. Insgesamt verteilen sich die Fehler vor der Korrektur auf 33 Assemblierungen, nach der Korrektur auf sieben in Tabelle A-5 und Tabelle A-6 im Anhang ist die genaue Anzahl der Fehler pro Konsensus-Sequenz dargestellt). Auffällig ist, dass bis auf eine Lücke, eine relativ große Anzahl Mismatches nach der Korrektur vorhanden ist.

Die häufigsten Fehler waren Mismatches, gefolgt von Lücken und Ns. Es konnten alle Ns, die vor allem Resultat schlechter Signalstärken waren, vollkommen entfernt werden. Dies war nur möglich, weil die Peaks noch deutlich erkennbar und sehr gut aufgelöst im Trace abgebildet wurden (Abbildung 6-10). Der Grund für das Einfügen der Ns liegt beim Base-Calling Algorithmus. Ein N wird dann eingefügt, wenn zu einem vorhergesagten Peak kein beobachteter Peak gefunden werden kann (vgl. 3.3.2) [Ewing, 1998 a].

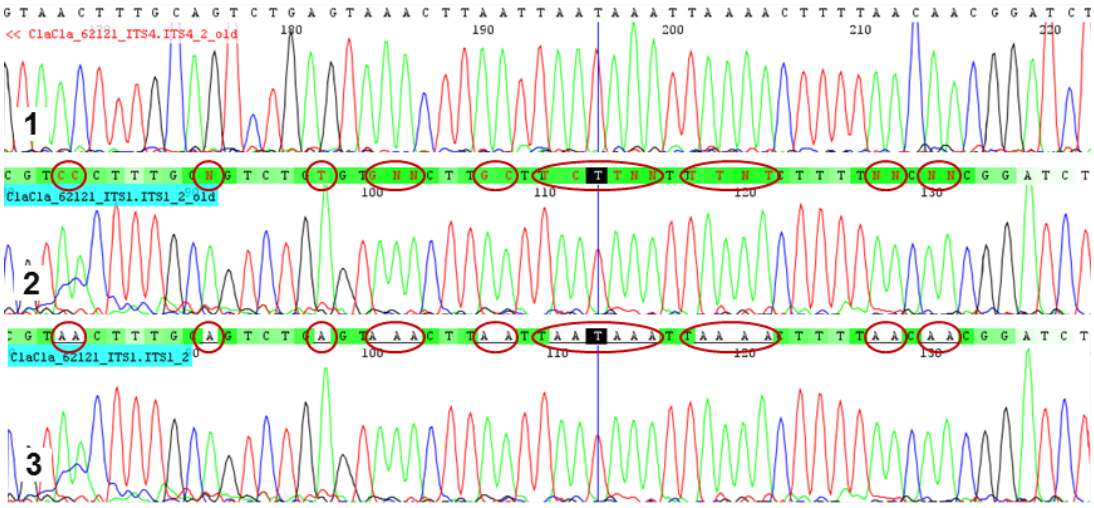

Abbildung 6-10: Einfluss schlechter Signalstärke auf die Konsensus-Sequenz

Das Beispiel zeigt die Assemblierung von *ClaCla* in der Elektropherogramm-Ansicht und den Einfluss von schwachen Signalstärken auf die Richtigkeit der Konsensus-Sequenz. Unter Nummer 1 ist der qualitativ gute Trace *ClaCla R* (ClaCla_62121_ITS4.ITS4_2) als Reverse+ Komplement dargestellt. Trace 2 und 3 zeigt *ClaCla F* vor (2) und nach (3) der manuellen Korrektur. Grün hervorgehobene Basen bedeuten schlechte PHRED-Werte (dunkelgrün <10, hellgrün 10-30) und rote Buchstaben Mismatches in der Konsensus-Sequenz. Aufgrund der schwachen Signalstärke von *ClaCla F* wurden Ns (z.B. Trace 2 Base 130) oder falsche Basen (z.B. Trace 2 Base 110) eingefügt obwohl der Trace klare Peaks besitzt. Die Signalschwäche ist in der Abbildung nicht zu erkennen, da die Höhe der Peaks in Trace 2 und 3 an die Höhe von Trace 1 angepasst worden ist. Durch manuelle Korrektur (unterstrichene Basen in Trace 3, rote Einrahmung) konnten diese Fehler behoben und eine eindeutige Konsensus-Sequenz hergestellt werden.

Ebenfalls konnten, bis auf eine, alle Lücken entfernt werden. Sie wurden vor allem am Anfang oder Ende eines Trace beobachtet, wo die Auflösung der Peaks noch nicht optimal ist. Dabei werden oft zwei gleiche aufeinanderfolgende Basen in einem Trace als nur eine Base erkannt (vgl. Abbildung 6-11). Grund dafür könnte eine zu schnelle oder zu langsame Wanderung der DNA-Fragmente während der Gelelektrophorese und die damit verbundenen Peakverschiebungen bzw. – Überlagerung sein (vgl. Kapitel 3.2.4).

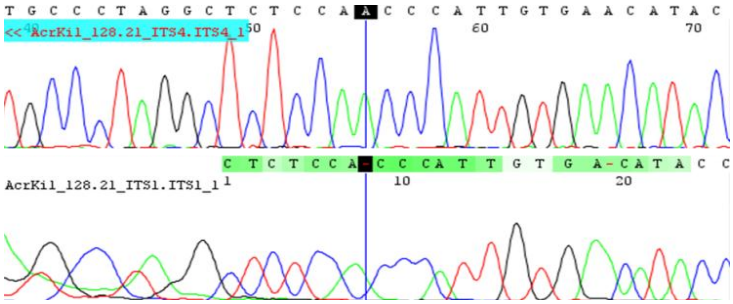

Abbildung 6-11: Auswirkung einer schlechten Auflösung zu Beginn eines Trace auf den Contig

Dargestellt ist die Assemblierung von *AcrKil* in der Trace-Ansicht. Die Position der Lücke im unteren Trace (schwarz hervorgehoben) befindet sich in dem schlecht aufgelösten Bereichs am Anfang. Der zweite A-Peak ist hier nicht erkennbar im Gegensatz zum deutlichen Peak bei dem oberen Trace.

Die Beseitigung der Mismatches dagegen, war weniger gut möglich. Die 21 verbleibenden Mismatches verteilen sich auf sieben *Contigs* (vgl. Tabelle A-6) und konnten durch Irregularitäten, die in den Traces auftraten, nicht beseitigt werden. Bei *AbsGla* verhinderte ein Primer-Fehler die genaue Zuordnung der Peaks ähnlich wie bei *AcrMur* (durch schlechte Auflösung bedingte Sekundärpeaks), *EngAlb* (Kontamination), *PenGri* (Signalschwäche) und *SynRac* (Doppelpeaks durch Chimäre). Bei *FusVer* ist eine Deletion, bzw. Verschmelzung zweier Peaks am Ende des Trace und bei *RhiOry* ein kleiner *Dye-Blob* Ursache für nicht korrigierbare Mismatches.

Betrachtet man die Ursachen für Fehler im Gesamten (Abbildung 6-12), so fällt auf, dass Signalschwäche und schlechte Auflösung (an den Enden des Traces) die Hauptursache für Fehler bei der Assemblierung darstellen. *Dye-Blobs* haben ebenfalls einen größeren Einfluss. Dies ist damit zu erklären, dass dadurch die Peaks mehrdeutig interpretiert werden können und das Base-Calling Programm durch z.B. einen nur geringfügig höheren „Blob-Peak“ eine falsche Base einfügt, obwohl der Trace unter dem *Dye Blob* eindeutig ist (vgl. Abbildung 6-2 ‚rechts).

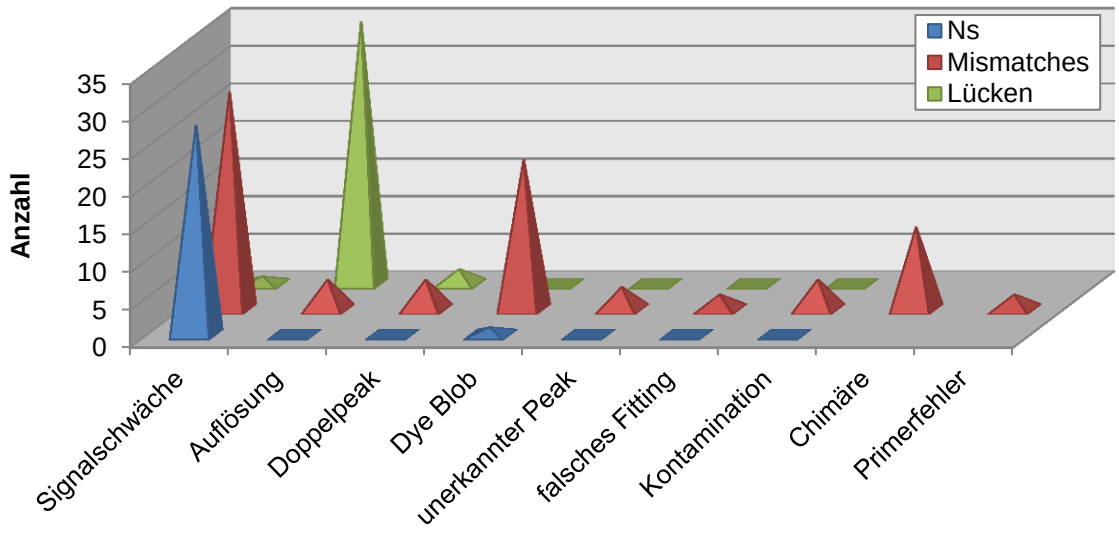

Abbildung 6-12: Ursache für Mismatches, Lücken und Ns und deren Häufigkeit

Im Diagramm werden die möglichen Ursachen für Fehler in Assemblierungen und deren Verteilung dargestellt. Es sind vor allem Signalschwächen für Mismatches und Ns verantwortlich. Eine schwache Auflösung der Peaks am Ende oder Anfang eines Trace führt eher zu Lücken. Einen relativ großen Einfluss auf das Einfügen von Mismatches haben auch *Dye-Blobs*.

6.5 Sequenzen in BOLD

6.5.1 *Dateninhalt*

Die Analyse des Inhalts von BOLD ergab, dass von insgesamt 25 Projekten mit Pilz-Kontext lediglich acht Projekte Sequenzen von Zielspezies[1] enthalten. Diese stammen nicht nur aus dem ITS-Bereich, sondern auch von dem COI-Gen, welches bereits als Barcode-Standard für Tiere akzeptiert wurde (vgl. Kapitel 4.3.1) und der großen (28S) und kleinen (18S) Untereinheit des Ribosoms (vgl. Kapitel. 4.3.2). In Abbildung 6-13 ist die Anzahl der Projekte in Abhängigkeit der Target-Sequenzen[2] dargestellt. Es sind vor allem ITS- und COI-Sequenzen in der Datenbank vertreten.

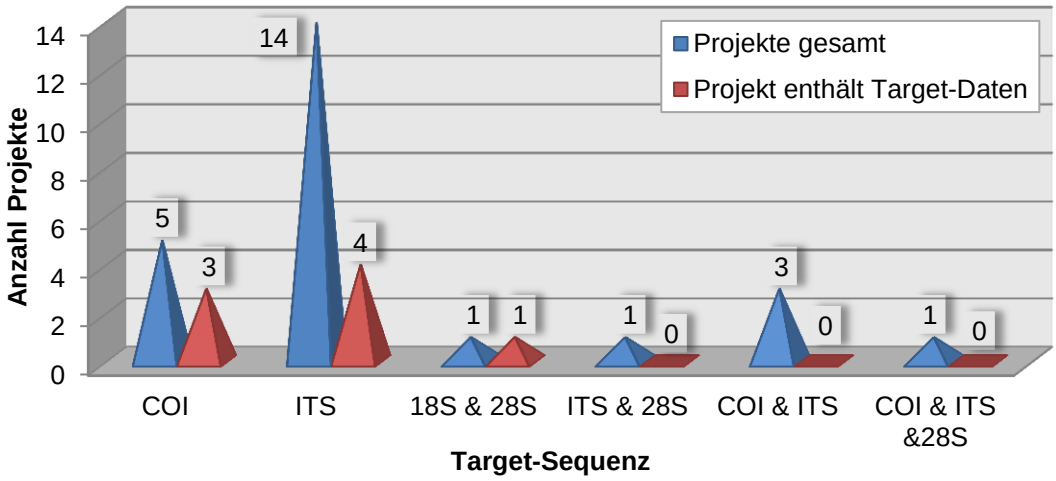

Abbildung 6-13: Verteilung der Target auf die Projekte in BOLD

Das Diagramm stellt die Verteilung der Projekte in BOLD auf die jeweiligen Target-Sequenzen dar. Es gibt Projekte, die jeweils nur einen Sequenz-Typ (nur COI, nur ITS) beinhalten und gemischte Projekte. Insgesamt sind 25 Projekte (blau), die Pilz-Sequenzen enthalten in BOLD aufgelistet. Jedoch nur acht (rot) davon besitzen Sequenzen, die mit unseren Zielspezies korrelieren.

Alle 24 Projekte zusammen enthalten 791 COI-Sequenzen, 3042 ITS-Sequenzen, 417 Sequenzen von 28S und eine aus 18S. Betrachtet man jedoch die Sequenzen, die mit den Zielspezies übereinstimmen, so ist theoretisch nur ein Bruchteil davon für Analysen nutzbar. Die acht Projekte enthalten nur 53 COI -, 59 IST - und eine 28S - Sequenzen. Des Weiteren konnte fast zu jeder Sequenz ein oder mehrere EPG

[1] Als Zielspezies werden die in Tabelle A-1 ausgelisteten Pilz-Spezies bezeichnet.
[2] Als Target werden die DNA-Sequenzen bezeichnet, die in der BOLD abgespeichert sind.

gefunden werden (92 COI, 120 ITS). Die Sequenzen und EPG verteilen sich auf insgesamt 24 der 100 Zielspezies Eine Übersicht ist in Tabelle 6-5 gegeben. Zu welcher Zielspezies in BOLD Sequenzen vorliegen und um welches Target es sich dabei handelt ist im Anhang in Tabelle A-1 dargestellt.

Tabelle 6-5: Statistik des Dateninhaltes in BOLD

Die Analyse des Dateninhalts mit Pilz-Kontext ergab die in der Tabelle dargestellten Ergebnisse. Insgesamt sind für Pilze vier verschiedene Target-Sequenzen abgespeichert. Dabei handelt es sich um den Barcode-Standard COI, den ITS-Bereich und deren flankierende Gene für die 18S- und 28S- Einheit des Ribosoms. Der verwendbare Datensatz (Sequenzen die von den genannten Zielspezies in Tabelle A-1 stammen) stellt jedoch nur einen Bruchteil der Daten dar, die in BOLD zu Pilzen enthalten sind.

Anzahl der …	Zielregionen			
	COI	ITS	28S	18S
Sequenzen gesamt	791	3042	417	1
Sequenzen von Zielspezies	53	59	1	1
EPG gesamt	1111	4011	0	0
EPG von Zielspezies	92	120	0	0
Spezies in BOLD gesamt	360	1215	45	1
Zielspezies in BOLD	17	9	1	1
Ziespezies insgesamt abgedeckt	24 von 100			

Da für die Entwicklung einer Diagnostik von humanpathogenen Pilzen bisher jedoch nur ITS –Sequenzen bzw. die gesamte ribosomale DNA als Ausgangsdaten genutzt wurden [Kropp, 2011], können praktisch nur die ITS-Sequenzen aus BOLD für direkte Vergleiche genutzt werden. Aus diesem Grund reduziert sich der Datenbestand weiter auf 59 Sequenzen und 120 Elektropherogramme in neun Zielspezies (vgl. Tabelle 6-5). Von diesen neun Spezies existieren wiederum nur sieben (mit 33 Sequenzen, 93 EPG), über die sequenzierte Daten vorliegen, für direkte Vergleiche zwischen den Sequenzen genutzt wurden. Die Projekte, aus denen die resultierenden Daten stammen, sind CEFI, GBF, WSF und PATE. Jedoch besitzt die Sequenz in PATE keine Elektropherogramme.

6.5.2 *Sequenzqualitäten im Vergleich*

Bei der Betrachtung der Elektropherogramme wurden einige Irregularitäten festgestellt. Am häufigsten traten *Dye-Blobs* (in 74 Traces) auf. Es wurden bis zu drei, unterschiedlich stark ausgeprägte *Dye-Blobs* pro Trace gezählt. Des Weiteren traten bei den „*failed*"-Sequenzen stark überlagerte oder verschwommene Peaks auf, was zu einer geringeren Qualität der einzelnen Basen führte. Bei ca. der Hälfte der EPG konnte zudem festgestellt werden, dass mit zunehmender Länge der Sequenz (und der in der Gelelektrophorese aufgetrennten DNA-Fragmente) die Peaks immer breiter wurden und asymmetrisch verschmierten (vgl. Abbildung 6-14).

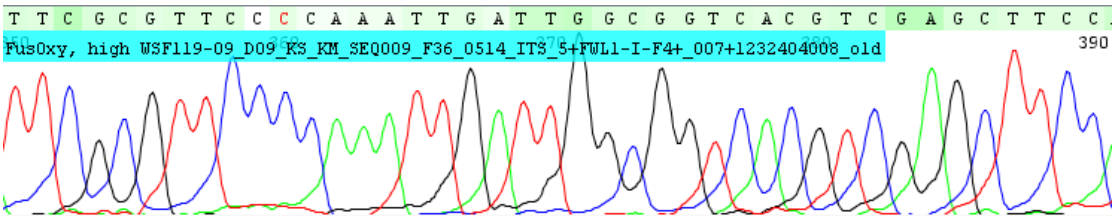

Abbildung 6-14: Unsymmetrische breite Peaks am Ende eines Trace (aus BOLD)

Der abgebildete Trace von *FusOxy* (Klassifizierung: high, BOLD-ID: WSF119-09 zeigt zunehmend asymmetrische, breiter werdende Peaks. Je größer die entsprechenden DNA-Fragmente bei der Gelelektrophorese sind, desto stärker kann dieser Effekt beobachtet werden.

Es könnte sich bei dieser Beobachtung um eine nur schwach ausgeprägte Peak-Verschiebung durch eine fehlerhafte Primer-Synthese (vgl. Kapitel 3.2.3) handeln. Da dieses Phänomen jedoch in fast der Hälfte der Traces auftrat und bei allen Klassifizierungen gleichermaßen vertreten war, kann dies ausgeschlossen werden. Eine weitere mögliche Ursache könnten auch unterschiedliche Wanderungsgeschwindigkeiten der DNA-Fragmente bei der Gelelektrophorese darstellen. Dies kann zum Beispiel durch Temperaturschwankungen des Gels hervorgerufen werden. [Lipshutz, 1994]

Mit Hilfe der Sequenzen aus BOLD sollte nun eine Validierung der Parameter, die in Kapitel 6.1.3 für die Klassifizierung von Sequenzen vorgestellt wurden, durchgeführt werden. Der Grund für die Verwendung der BOLD für diese Überprüfung ist die anerkannte und standardisierte Bewertung von Sequenzen. Dafür wurde zuerst die Einstufung der BOLD-Sequenzen ermittelt. Danach erfolgte die Anwendung der in Kapitel 6.1.2 und 6.1.3 beschriebenen Verfahren auf die BOLD-Sequenzen. Es ergab sich die in Tabelle 6-6 dargestellte Einteilung der Sequenzen auf die entsprechende Kategorie. Laut BOLD werden 62 Sequenzen als qualitativ sehr gut („high") und 17 als „medium"

eingestuft. Nach einer Bestimmung der PHRED-Werte, dem EndClipping und Anwendung der beschrieben Klassifizierungs-Kriterien waren jedoch nur 10 Sequenzen als „high“, dafür aber 64 als „medium“ eingestuft worden. Die Anzahl der „low“ – und „failed“- Sequenzen blieben jedoch relativ gleich. Insgesamt stimmten bei 28 Sequenzen die zugeordneten Kategorien bei BOLD und mit Hilfe der vorgestellten Kriterien (vgl. Kapitel 6.1.3) überein. Des Weiteren wurden 63 Sequenzen um ein Kriterium (z.B. in BOLD als „high“ bewertet, aber durch die Klassifizierungskriterien als „medium“) und zwei Sequenzen um zwei Kriterien herabgestuft (von „medium“ nach „failed“). Bei zwei weiteren Sequenzen konnte jedoch auch eine Erhöhung der Sequenzen auf eine qualitativ bessere Kategorie erzielt werden.

Tabelle 6-6: Klassifizierung der ITS-Traces von BOLD

Die Tabelle gibt wieder, welche Anzahl der Sequenzen aus BOLD in die jeweiligen Kategorien *„high“, „medium“, „low“* und *„failed“* eingeteilt wurden. Des Weiteren ist die Verteilung der Sequenzen nach Anwendung der in Kapitel 6.1.3 genannten Kriterien dargestellt. Während die Anzahl der *„low“* und *„failed“*-Sequenzen relativ gleich bleibt, variiert die Anzahl der *„high“* und *„medium“* Sequenzen deutlich..

Kategorie	high	medium	low	failed
Anzahl Sequenzen (nach BOLD)	62	17	8	6
Anzahl Sequenzen (nach eigenen Kriterien)	10	64	6	13

6.5.3 *Sequenzvergleiche*

Um Unterschiede zwischen den Konsensen der sequenzierten Daten und der Sequenzen aus BOLD sichtbar zu machen, wurde eine Alignierung mit Hilfe von BLAST durchgeführt *(„Align two or more sequences“*, Einstellung: *megablast)*. Falls mehrere BOLD-Sequenzen (nicht die Sequenzen der Traces, sondern die fertig assemblierten Sequenzen) für eine Spezies vorhanden waren, wurden diese vor dem Alignment assembliert. Dies betraf *FusOxy, FusSol* und *TriHar*.

Die Alignments *von AcrMur, FusOxy, FusSol* und *PenBre* ergaben eine gute Sequenzähnlichkeit und der Größe des übereinstimmenden Bereichs entsprach annähernd der der ITS-Region. Dagegen wurde bei *GeoCan* keine Ähnlichkeit gefunden und *GeoPan* sowie *TriHar* hatten nur einen kleinen ähnlichen Sequenzbereich. Dies deutete darauf hin, dass entweder die Sequenzen in BOLD oder die Sequenzierung fehlerhaft sind.

Um zu überprüfen welche Sequenz mehr Ähnlichkeit zu der entsprechenden Spezies besitzt, wurde eine BLAST-Suche gegen die „*Nucleotide collection*“ des NCBI durchgeführt.

Dabei lieferten die Sequenzen von *GeoCan* und *GeoPan*, die aus den selbst sequenzierten Datensatz stammten, den besten BLAST-Treffer, während die Sequenzen aus BOLD keinen Treffer (zur entsprechenden Spezies) hervorbrachten. Genau das Gegenteil ist bei *TriHar* der Fall, hier lieferte die BOLD-Sequenz gleich an erster Stelle einen Treffer. Die Ergebnisse der Sequenzvergleiche sind in Tabelle 6-7 zusammengefasst.

Tabelle 6-7: Ergebnisse des Alignments zwischen BOLD und sequenzierten Daten

Die Alignments wurden mittels BLAST („*Align two or more sequences*“, Einstellung: *megablast*) durchgeführt. Das beste Alignment mit 100 % Übereinstimmung liefert *PenBre*. Bei *GeoCan* wurden keine Ähnlichkeiten festgestellt. *GeoPan* und *TriHar* brachten ebenfalls nicht zufriedenstellende Ergebnisse.

Spezies	Identische Positionen	Lücken	E-Wert	Größe des Alignments	Besserer BLAST-Treffer
AcrMur	522 (98 %)	1 (0,2 %)	0,0	532	-
FusOxy	527 (99 %)	2 (0,4 %)	0,0	532	-
FusSol	478 (95 %)	14 (3 %)	0,0	510	-
GeoCan	Kein Übereinstimmungen gefunden				Sequenzierung
GeoPan	162 (96 %)	2 (1 %)	2e-77	169	Sequenzierung
PenBre	582 (100 %)	0	0,0	582	-
TriHar	227 (89 %)	10 (4 %)	2e-88	255	BOLD

6.6 Assemblierte Datenbanksequenzen

Es standen insgesamt 59 assemblierte Sequenzen von Zielspezies zur Verfügung. Die dafür verwendeten Sequenzen stammten größtenteils aus dem NCBI und der ArbSilva Datenbank. Sie beinhalten nicht nur die reinen ITS-Sequenzen sondern auch größere Teile des 18S- und 28S- Gens. Aufgrund der in Kapitel 3.4 geschilderten Probleme bei Pilzsequenzen in Datenbanken, sollten mit Hilfe der BOLD- und sequenzierten Daten die Qualität der Sequenzen überprüft werden. Die ursprüngliche Idee war, zu den verwendeten Datenbank - Sequenzen die entsprechenden Traces aus dem *TraceArchiv*[1] des NCBI zu nutzen, um die Sequenzen und deren Qualität anhand der Rohdaten bestimmen und überprüfen zu können. Anschließend sollten die standardisierten und qualitativ bewerteten Sequenzen aus BOLD für Vergleiche dienen. Da aber zu keiner Sequenz ein EPG im *TraceArchiv* vorliegt, mussten weitere Möglichkeiten zum Datenvergleich entwickelt werden: Nun wurden die erstellten Konsensus -Sequenzen der sequenzierten Daten sowie die Sequenzen aus BOLD[2] genutzt um mögliche Fehler in den assemblierten Datenbank- Sequenzen aufzudecken. Da nicht zu jeder assemblierten Sequenz Referenz-Sequenzen vorhanden war, reduziert sich die Anzahl der zu untersuchenden Daten auf 23 Assemblierungen (vgl. Tabelle A-1 im Anhang). Der Vergleich erfolgte mit dem Programm CodonCodeAligner, in dem die entsprechenden Sequenzen über die Funktion *Contig>Assemble* aliniert wurden. Anschließend wurde die Anzahl der Lücken und Mismatches im überlappenden Bereich ausgezählt. Die Ergebnisse sind im Anhang unter Tabelle A-8 aufgelistet. Es wurde zudem eine Unterscheidung zwischen der Anzahl Lücken und Mismatches im Anfangsbereich der Überlappung und dem Endbereich unternommen, die jeweils den ungefähren Positionen der ITS1 bzw. ITS2 Region entsprechen. Sie können durch einen mittleren Bereich von ca. 100 bis 300 Basen, der, auch bei einer sehr großen Anzahl auftretender Mismatches oder Lücken, immer fehlerfrei blieb, voneinander getrennt werden. Die Analyse ergab, dass 17 der assemblierten Sequenzen nur eine sehr geringe Fehlerzahl von null bis elf Mismatches bzw. Lücken aufwiesen. Oft traten solche Fehler am Anfang bzw. Ende der jeweiligen Referenz-Sequenz auf. Drei Assemblierungen (*AbsCor, AltAlt* und *WalSeb*) zeigten an 20 bis 26 Positionen Unterschiede und weitere drei weisen mehr als 50 Lücken bzw. Mismatches auf (*AbsGla, EngAlb* und *MucRac*).

[1] Das TracheArchiv ist eine Datenbank, in der Elektropherogramme zu Datenbankleinträgen abgespeichert werden können. Es ist unter der Web-Adresse http://www.ncbi.nlm.nih.gov/Traces/trace.cgi zu finden. Die Suche nach Trace-Daten wurden mit Hilfe der AC-Nummern der verwendeten Datenbank-Sequenzen durchgeführt (zum Beispiel: ACCESSION="AC245449")

[2] im Folgenden als Referenzsequenzen oder Referenz bezeichnet

7 Diskussion

7.1 Sequenzqualität

In dieser Arbeit wurden die sequenzierten Daten von 48 Zielspezies ausgewertet, um Kriterien für eine Qualitätsbewertung zu erarbeiten. Um mögliche Auswirkungen von Sequenzierungsfehlern auf die Sequenzen überprüfen zu können, wurden zuerst alle vorliegenden 102 Elektropherogramme auf Irregularitäten untersucht (vgl. Kapitel 6.2). Anschließend erfolgte die Konvertierung der Traces in DNA-Sequenzen (vgl. Kapitel 6.3) sowie deren Klassifizierung. Der letzte Schritt beinhaltete die Assemblierung der Forward- und Reverse-Sequenzen, um Sequenz-Unterschiede zu detektieren und manuell zu korrigieren (vgl. Kapitel 6.4).
Es konnten einige Irregularitäten entdeckt werden (vgl. Abbildung 6-7). Die möglichen Ursachen wurden schon in Kapitel 3.2 erläutert und werden daher hier nicht diskutiert. Abhängig von Art und Ausprägung einer Irregularität konnten unterschiedliche Auswirkungen auf die Klassifizierung einer Sequenz und der Assemblierung festgestellt werden (vgl. Tabelle A-3, Tabelle A-4, Tabelle A-5 und Tabelle A-6):

1 *DyeBlob*:

Er führt zu lokal sehr schlechten PHRED-Werten, der im Durchschnitt fünf bis zehn Basen, welche an der Stelle des *Dye-Blobs* vom Base-Calling Algorithmus geschlussfolgert wurden. Diese liegen zwischen PHRED 15 und PHRED 25 am Rand des *Blobs*, und unter PHRED 10 im mittleren Bereich des *Blobs*[1]. Angenommen eine Sequenz besitzt eine Länge von 500 Basen und ein *Dye-Blob* verursacht zehn Basen mit PHRED-Werte unter 20, dann kann diese Sequenz schon nicht mehr als „*high*“ eingestuft werden, weil das Verhältnis von Ns (PHRED < 20) zu Gesamtlänge der Sequenz schon 4 % beträgt (vgl. Kapitel 6.1.3). Daraus ergibt sich die Klassifizierung der 15 Sequenzen, die nur einen Dye-Blob als Irregularität aufwiesen: neun *„medium“*- und sechs *„high“*-Sequenzen.

Vorteil von *Dye-Blobs* gegenüber anderer Irregularitäten ist die hohe Wahrscheinlichkeit, dass die Sequenz im Bereich des *Dye-Blobs* trotzdem gut abgelesen

[1] Die Daten sind hier nicht gezeigt, können aber auf der beigelegten CD im Ordner „Excel-Tabellen“ Datei „Sequenzqualitäten“ nachvollzogen werden.

werden kann, da oft der Verlauf der Peaks nicht beeinflusst wird (vgl. Abbildung 6-2, rechtes Bild). Dies wird vor allem bei der manuellen Überprüfung der erstellten Konsensen aus Forward- und Reverse-Sequenz deutlich. Es konnten alle 20 Mismatches und das eine N, welche aufgrund von Dye-Blobs entstanden sind, entfernt werden.

2 ***Sehr schwache und schwache Signalstärken****:*

Wie in Kapitel 6.3 bereits erwähnt, führen sehr schwache Signalstärken vorwiegend zu sehr kurzen (< 10 bp) bzw. kurzen (300-450 bp) Sequenzen. Insgesamt 13 Sequenzen der 17 stark signalschwachen Traces sind nur 1 bis 10 bp lang. Die Ursache dafür liegt beim EndClipping. Laut den in Kapitel 6.1.2 beschriebenen Parametern wird eine Sequenz so lange gekürzt bis in einem Fenster von 20 (50 am Ende der Sequenz) die Fehlerwahrscheinlichkeit im Durchschnitt unter 0,01 (größer PHRED 20) liegt. Da aber für Basen in signalschwachen Traces in der Regel nur sehr kleine PHRED-Werte berechnet werden, können diese nicht erreicht werden und die Sequenz wird immer weiter gekürzt. Bis auf *PenGri R*, der noch relativ deutliche Signale im Trace zeigt und somit qualitativ bessere PHRED-Werte enthält, sind alle stark signalschwachen Traces als *„failed“* eingestuft worden.

Ein wenig besser wurden die sieben, nur als signalschwach bezeichneten, Sequenzen klassifiziert. Sie besitzen gute Sequenzlängen ab 450 bp und sind, insofern nur die signalschwäche als einzige Irregularität besteht, als *„medium“* eingestuft worden (ansonsten *„failed“*).

Die Assemblierungen der Sequenzen mit (sehr) schwachen Signalstärken sind entweder fehlgeschlagen oder besaßen eine große Anzahl an Fehlern im Contig (vgl. Tabelle A-5, *ClaCla, FusSol, PenBre, PenGri*). Dies lässt sich auf ungenaues Base-Calling durch starkes Rauschen, undeutlichen Peaks oder Sekundärpeaks zurückführen. Jedoch hat die manuelle Korrektur gezeigt, dass es trotz der schlechten Traces möglich ist, die Fehler zu beseitigen. Es konnten lediglich zwei Mismatches bei der Konsensus von *PenGri* nicht berichtigt werden (vgl. Tabelle A-6).

3 ***Fehlerhafte Primer-Synthese, Polymerase-Slippage, Chimäre und Kontamination:***

Zu jeder dieser Irregularitäten liegen zwei Traces vor. Dabei besitzt ein Trace einer stärkeren Ausprägung als der andere. Zum Beispiel tritt die Chimäre einmal ab Base 430 und einmal schon ab Base 130 auf und der Slippage entsteht einmal gleich am Anfang des Trace und einmal erst ab Base 130. Bei der Kontamination mit Fremd-DNA und dem Primer-Fehler variieren die Höhen der Sekundärpeaks

unterschiedlich stark. Das führt dazu, dass jeweils der Trace mit der schwächeren Ausprägung als *„medium“* gewertet wurde. Denn obwohl ein Fehler vorliegt, konnten noch große Sequenzbereiche, relativ guter Qualität isoliert werden. Die anderen Traces wurden dementsprechend schlechter eingestuft und erhielten den Status *„failed“* bzw. *„low“* (bei *EngAlb R*, Kontamination).

Da eine der beiden Sequenzen eher schlechtere Qualität besitzt und die vier Irregularitäten durch Sekundärpeaks bzw. komplett verschobenen Peaks beim Slippage gekennzeichnet sind, ist die Wahrscheinlichkeit einer falsch eingefügten Base beim Base-Calling relativ hoch. Dies führt bei der Assemblierung zu einer ähnlichen Anzahl von Lücken, Mismatches und Ns von den (stark) signalschwachen Traces. Der Slippage führte dazu, dass keine Konsensus-Sequenz gefunden werden konnte. Durch die manuelle Berichtigung der drei Konsensen von *AbsGla* (Primer-Fehler), *EngAlb (*Kontamination) und *SynRac* (Chimäre) konnten jedoch nur ein geringer Teil der Fehler korrigiert werden. Dies liegt daran, dass durch die teilweise hohen Sekundärpeaks nicht eindeutig entschieden werden kann, welcher von beiden der richtige ist. Es könnte sich schließlich auch um einen SNP handelt. Um solche Fehler zu bewerten hilft nur eine erneute Sequenzierung.

Aus den vorgestellten Beobachtungen kann man folgende Schlussfolgerungen ziehen:
Die Klassifizierung ist in der Lage verschieden starke Ausprägungen von Irregularitäten zu erkennen und stuft die Sequenzen, die wahrscheinlich Fehler beinhalten herab (siehe Punkt 2 und 3). Des Weiteren haben Dye-Blobs, die einen eher geringen Einfluss auf die Richtigkeit der Sequenz besitzen, nicht dazu geführt, dass eine Sequenz als *„low“* oder *„failed“* eingestuft worden ist. Im Gegenzug wurde jedoch auch keine Sequenz mit z.B. einem stark signalschwachen Trace als „medium“ oder „high“ bezeichnet. Die einzige Ausnahme hierbei bildet die Sequenz von *PenGri R*, die den Status „medium“ erhielt. Jedoch zeigt der Trace, trotz der geringen Signalstärke, klare Peaks. Diese Erkenntnisse lassen den Schluss zu, dass die Kriterien sensitiv genug sind, um Sequenzen nach ihrer Qualität einzuteilen. Das vierstufige System ist ebenfalls gut dazu geeignet schnell einen Überblick über eine Datensammlung zu bekommen und gute Sequenzen auszuwählen.
Was trotz der guten Ergebnisse der Klassifizierung noch optimiert werden könnte, sind die Parameter für das EndClipping. Da es sich um eine mittelwertbasierte Methode handelt, befinden sich am Anfang und Ende einer Sequenz oft noch einige (ca. 2 bis 6) Basen mit sehr kleinen PHRED-Werten. Diese könnten zum Beispiel das Ergebnis

einer Klassifizierung beeinflussen (Parameter „Anteil Ns“). Wenn diese durch zum Beispiel zweiphasiges[1] EndClipping entfernt werden könnten, würde man zum einen eine Abstufung von „high“ zu „medium“ nur Aufgrund dieser schlechten Basen verhindern. Zum anderen würde ein solcher Schritt möglichen Fehlern bei der Assemblierung von Sequenzen vorbeugen. Denn Lücken und Mismatches treten bei den Konsensen sehr häufig an den Stellen auftreten, an denen sich die Sequenzen beginnen zu überlappen (vgl. Abbildung 7-1). Dies würde das Contig in seiner Qualität (weniger Fehler) steigern und die Zeit, die zur manuellen Korrektur benötigt wird verringern.

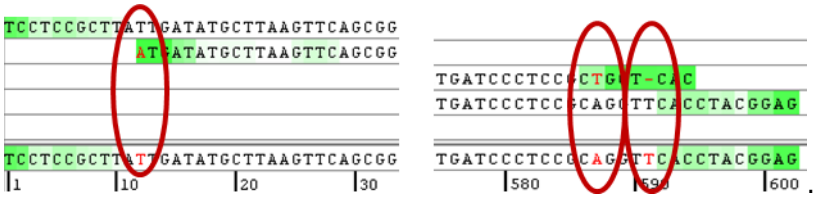

Abbildung 7-1: Assemblierungs-Fehler am Anfang/Ende eines überlappenden Bereichs

Die Abbildung zeigt, welchen Einfluss qualitativ schlechte Basen, die nicht durch das EndClipping entfernt wurden, auf eine Assemblierung haben. Es entsteht ein Mismatch (linkes Bild; rechtes Bild, linkes Oval) und eine Lücke (rechtes Bild, rechts oval).
Die grün hervorgehobenen Basen haben einen PHRED-Wert < 20.

Insgesamt kann man jedoch die Qualität der Konsensus-Sequenzen als sehr gut einstufen. Zum einen wird dies natürlich durch die hohe Anzahl an „high“ und „medium“-Sequenzen im Datensatz verursacht. Zum anderen hat die manuelle Kontrolle jedoch sehr zur Verminderung von Lücken, Mismatches und Ns beigetragen (vgl. Tabelle A-5 und Tabelle A-6). Die nach der Korrektur noch vorhandenen Mismatches könnten noch durch eine erneute Sequenzierung der ITS-Bereiche des entsprechenden Organismus und anschließenden Assemblierung mit den schon vorhandenen Sequenzen beseitigt werden.

Insgesamt kann man die Datenqualität der vorliegenden Sequenzen als sehr gut bezeichnen. Da fast alle Contigs nach der manuellen Korrektur keine Fehler mehr besaßen bzw. die Anzahl der Mismatches sehr gering war, können diese Daten auch im weiteren Verlauf des Projektes genutzt werden. Ausnahme hierbei stellt *SynRac* dar. Durch das Auftreten chimärer Sequenzen und der hohen Anzahl an Mismatches sollte diese Sequenz aus dem Datensatz ausgeschlossen werden.

[1] Damit ist die zweimalige Anwendung des EndClppings auf einen Datensatz gemeint. Die erste Phase sollte so gewählt sein, dass der Großteil der schlechten Sequenzabschnitte entfernt wurde. Der zweite Schritt dient dann der Entfernung restlicher qualitativ schlechter Basen am Anfang und Ende der Sequenz.

7.2 BOLD Daten

Betrachtet man die gesamten, zu Pilzen enthaltenen Daten in BOLD, so hat man eine ausreichende Datenmenge für zum Beispiel phylogenetische Analysen vorliegen. Des Weiteren mag vorteilhaft sein, dass sowohl Sequenzen des COI – Gens als auch der ribosomale DNA in der Datenbank abgespeichert sind. Denn beide Sequenzen stellen mögliche Kandidaten für Barcode-Regionen dar (vgl. Kapitel 4.3.2).
Für die Zielspezies dieses Projektes ist jedoch nur ein Bruchteil der benötigten Daten in BOLD enthalten.
Sie sollten ursprünglich dazu dienen die sequenzierten Daten und die vorliegenden assemblierten Daten aus den Datenbanken zu validieren, um klären zu können wie gut sie für die die Entwicklung artspezifischer Oligo-Nukleotide zur Identifikation humanpathogener Pilze geeignet sind. Des Weiteren soll überprüft werden, welche Ergebnisse die vorgestellte Klassifizierung bei BOLD-Sequenzen erzeugt.
Insgesamt sind jedoch nur sieben Spezies in BOLD von denen sequenzierte Daten vorliegen und drei Spezies von denen Datenbank-Sequenzen vorhanden. Da jedoch zu einigen der sieben Spezies mehr als eine Sequenz und mehrere Traces vorhanden waren, wurde trotzdem ein Vergleich durchgeführt.

7.2.1 *Vergleich mit sequenzierten Daten*

Zu Anfang wurden, wie bei den Sequenzierten Daten alle Elektropherogramme aus BOLD nach Irregularitäten untersucht (vgl. Kapitel 6.5.2). Das Hauptproblem sind *Dye-Blobs*, wie bei den sequenzierten Daten, nur dass in den BOLD-Traces oft zwei Blobs in einer Sequenz auftreten. Dies hat zur Folge, dass mehrere Basen einen schlechten PHRED-Wert besitzen und dadurch die Sequenzen, die in BOLD als „high“ eingestuft worden sind, nun den Status „medium“ besitzen, obwohl diese keine weiteren Irregularitäten besitzen. Weiterhin wird in der BOLD -Datenbank nicht das beschriebene Zusatzkriterium zur Klassifikation von Sequenzen verwendet. Außerdem wird im BOLD-Datenstandard eine qualitativ hohe Base mit PHRED-Werte > 20 definiert. In der Klassifikation liegt der Wert dafür jedoch bei PHRED 30. Das hat zur Folge, dass in BOLD der Mittelwert aller guten Basen (PHRED > 30), der bei rund PHRED 47 liegt, niedriger ist als der der sequenzierten Daten (51,3), weil dadurch auch Sequenzen mit niedrigeren PHRED-Werten in die Datenbank aufgenommen werden. Die beobachteten verschwommenen Peaks am Ende eines Trace haben

jedoch keinen Einfluss auf die Sequenzqualität und könnten, da sie in fast jedem Trace auftreten, auf die verwendete Sequenzier-Methodik zurückzuführen sein.
Um diese unterschiedlichen Ergebnisse der Klassifizierung eindeutig bewerten zu können, sollten jedoch noch weitere Analysen durchgeführt werden. Zum Beispiel könnten die niedrig bewerteten Basen, an den Dye-Blob Positionen per Hand korrigiert werden. Dadurch könnte jeder Base ein maximaler PHRED-Wert zugeordnet und erneute Klassifizierungen durchgeführt werden. Stimmt das Ergebnis besser mit der BOLD-Klassifizierung überein, sollten weitere Kriterien entwickelt werden, die noch sensitiver reagieren. Ist dies nicht der Fall, so sollten mögliche Fehlerquellen in den Klassifizierungs-Kriterien abgeklärt und behoben werden

Der anschließende Vergleich der Sequenzen der sieben Spezies (vgl. Kapitel 6.5.3) lieferte sowohl gute als auch schlechte Ergebnisse (vgl. Tabelle 6-7). Als gutes Ergebnis werden die vier fast fehlerfreien Alignments bewertet. Sie sind nahezu identisch und weisen einen E-Wert von null auf. Das bedeutet, das gefundene Alignment ist signifikant. Die eingefügten Lücken oder Mismatches (ergeben sich aus Größe des Alignments-(Lücken + identische Positionen)) entstehen durch die intra-spezifische Varianzen und sind bei dem Vergleichen von Sequenzen verschiedener Organismen einer Spezies zu erwarten. Weniger gut sind dagegen die verbleibenden drei Alignments, die nur eine unzureichende Ähnlichkeit aufweisen. Die Datenbank-suche ergab zudem, dass bei *GeoPan* und *GeoCan* die Contigs der Sequenzierung ähnlicher zu der wirklichen Sequenz der Spezies sind als die Sequenzen aus BOLD. Dies würde bedeutet, dass die Daten aus BOLD fehlerbehaftet sind und nicht aus der angegebenen Spezies stammen. Solche Fehler können passieren, wenn zum Beispiel bei der Kultivierung der Pilze keine Reinkultur vorliegt und für die Sequenzierung genau die falschen Organismen von dem Nährmedium genutzt werden oder Sequenzier-Ansätze vertauscht wurden.
Zusammenfassend betrachtet, kann man anhand von sieben Sequenzvergleichen keine Aussage über die gesamte Datenqualität getroffen werden. Es ist sicher, dass die vier sehr guten Alignments der BOLD und sequenzierten Daten stellvertretend für eine hohe Sequenzqualität stehen. Wie sich diese jedoch im restlichen Datenbestand verhält kann nicht beurteilt werden.

7.3 Beurteilung der assemblierten Datenbanksequenzen

Es konnten alle 23 Datenbanksequenzen mit den jeweiligen Referenzen aligniert werden. Bei der Auswertung der Alignierung konnten unterschiedlich viele Lücken und Mismatches pro Sequenz festgestellt werden. Dabei viel auf, dass die Sequenzen mit mehreren Fehlern in der Mitte des alignierten Bereichs keinerlei Unterschiede aufwiesen. Dieser Bereich ist ca. 100 bis 300 bp lang. Da es sich bei den untersuchten Sequenzen um den ITS-Bereich der ribosomalen DNA handelte und dieser die konservierte 5,8S-Sequenz beinhaltete, wird vermutet, dass die 100 – 300 bp diese konservierte Region darstellen (vgl. Abbildung 7-2). Dann sind jeweils vor den konservierten Bereich die ITS1 und danach die ITS2 –Region lokalisiert, wo die Lücken und Mismatches auftreten. Um gleichzeitig eine Aussage über mögliche Unterschiede in der Anzahl der Fehler in der ITS1 und ITS2-Region zu erhalten, wurden diese getrennt betrachtet.

Die drei Alignierungen mit den meisten Fehlern sind *AbsGla* (81), *MucRac* (50) und *EngAlb* (112). Eine mögliche Erklärung für die starken Unterschiede wäre, dass in den Assemblierungen der Referenzdaten Fehler existieren, die bisher nicht erkannt wurden. Dagegen spricht jedoch, dass in dem Bereich der vermuteten 5,8S-Region keinerlei Lücken oder Mismatches auftreten, was man bei einer fehlerhaften Referenz - Sequenz aber durchaus erwarten würde. Es wäre also möglich dass zwischen den Sequenzen dieser Spezies eine hohe intraspezifische Varianz Folge der hohen Fehleranzahl ist. Betrachtet man die Längen der jeweiligen überlappenden Bereiche aller Alinierungen, dann stellt man fest, dass diese der Länge des ITS-Bereichs entsprechen.
Die Anzahl der Lücken im ITS1 – und ITS2- Bereich unterschieden sich nur minimal (Lücken ITS1: 51, ITS2: 55; Mismatches ITS1: 147, ITS2: 122). Daraus lässt sich ableiten, dass die beiden Regionen ungefähr gleiche Variabilität besitzen.
Zusammenfassend kann man sagen, dass die untersuchten Datenbanksequenzen von guter Qualität sind. Sie besitzen keine großen nicht übereinstimmenden Bereiche, wie sie bei dem Vergleich der sequenzierten Daten mit BOLD-Sequenzen auftraten. Aufgrund der konservierten Region im Mittelteil des alignierten Bereichs, die im Alignment keine Lücken oder Mismatches enthält, kann man auch bei den Sequenzen mit hohen Anteil an Fehlern eine hohe Sequenzqualität vermuten.

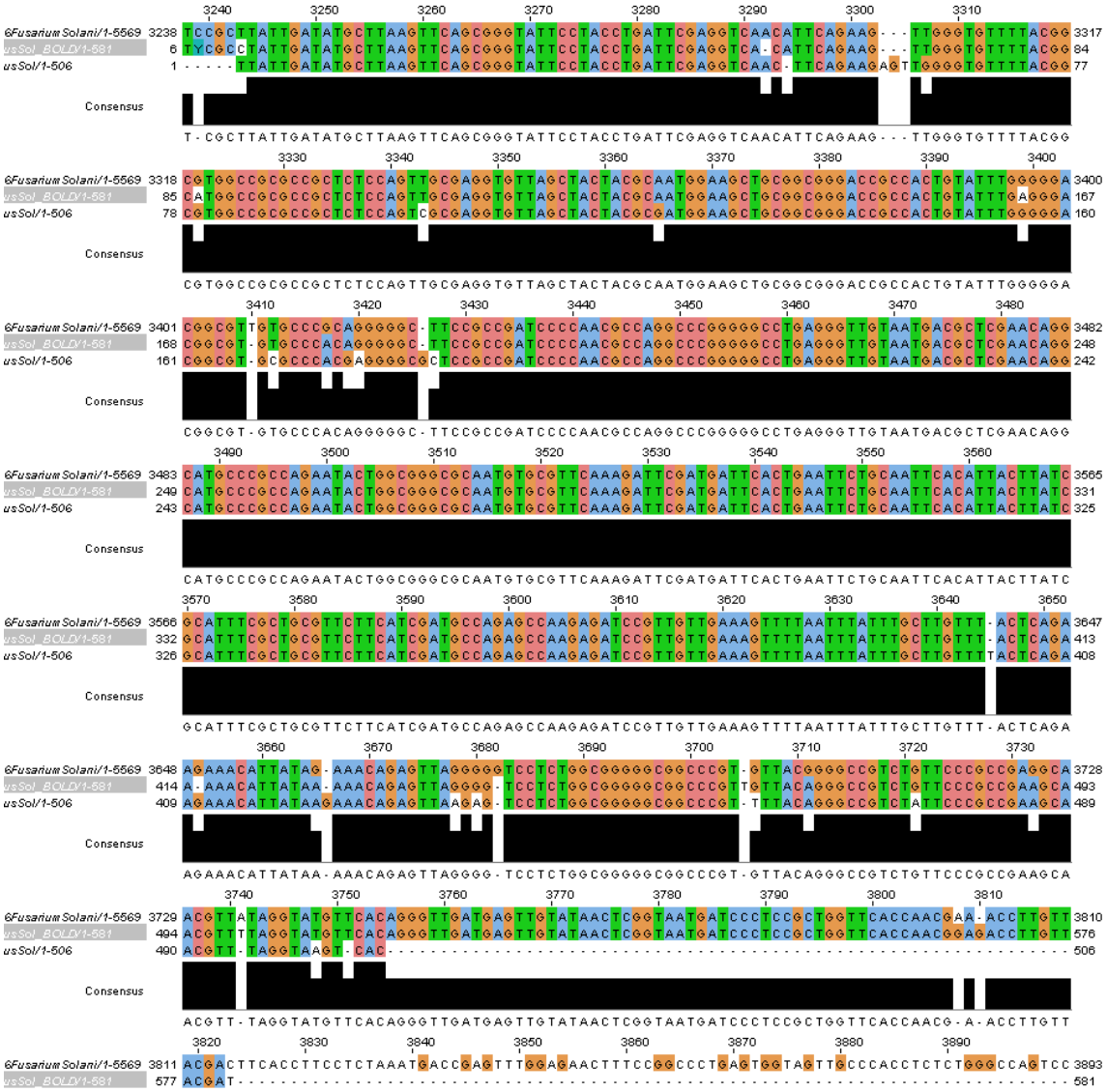

Abbildung 7-2: Konservierte Region bei Überprüfung der Datenbanksequenzen

Die Abbildung zeigt ein Alignment der vorliegenden Referenz-Sequenzen aus BOLD und vom sequenzierten Datensatz mit der assemblierten Datenbank Sequenz von FusSol. Es ist trotz der höheren Anzahl von Lücken und Mismatches ein konservierter Bereich in der Mittel der Sequenz erkennbar (Base 3427-3645). Diese Region ist vermutlich die konservierte 5,8S.Sequenz der ribsomalen DNA.

8 Ausblick

Die Identifikation humanpathogener Pilze auf Grundlage von DNA-Sequenzen setzt eine hohe Datenqualität und eine ausreichend große Datensammlung voraus. Die vorgestellten sequenzanalystischen Schritte sowie die Kriterien zur Klassifizierung von Sequenzen ausgehend von ihren Rohdaten (EPG) stellen eine gute Basis für das Gelingen eines solches Projektes dar. Die Überprüfung der bereits vorliegenden Sequenzen ergab, dass diese den qualitativen Anforderungen entsprechen. Der Datenpool konnte durch die Sequenzierung von ITS-Sequenzen, deren Qualität zuerst bestimmt wurde, erweitert werden.

Weiterhin stellt die Analyse dieser Sequenzen einen wichtigen Punkt in der vorgestellten Strategie dar, dabei besitzt die Überprüfung der Datenqualität einen hohen Stellenwert besitzt.

Diese Prozesse können in weiteren Arbeitsschritten noch verfeinert, standardisiert und zusammengefasst werden, damit sie in die Strategie eingearbeitet werden können. Als großes Ziel soll hierbei die Entwicklung einer Software genannt werden, die die vorgestellte Strategie wiedergibt und mit der verschiedenste biologische Aufgaben-stellungen bearbeitet werden können.

Literatur

[Achaz, 2008]
Achaz, Guillaume: *Testing for Neutrality in Samples With Sequencing Errors.* In: Genetics. - Bethesda: Genetics Society of America- 179(Juli 2008)3, S. 1409-1424

[Altschul, 1990]
Altschul, Stephen F.; Gish, Warren; et. al.: Basic Local Alignment Search Tool. In: J. Mol. biol. – Kidlington: Elsevier B.V. – 215(1990)3, S. 403-410

[Begerow, 2010]
Begerow, Dominik; Nilsson, Henrik; et. al.: *Current state and perspectives of fungal DNA barcoding and rapid identification procedures*. In: Appl. Microbiol. Biotechnol. – Berlin Heidelberg: Springer-Verlag – 87(2010)1 , S. 99 - 108

[Brettschneider, 2011]
Brettschneider, Janine: *Generierung von Oligonukleotiden zur Identifikation von Candida-Arten durch die Entwicklung eines darauf angepassten Algorithmus*‘. – 2011. – Mittweida, Hochschule Mittweida, Mathematik/Naturwissenschaft/Informatik, Bachelor-Arbeit, 2011

[BOL Data Portal, 2011]
Trizna, Mike <triznam@si.edu>: *The Barcode of Life Data Portal*. URL: http://bol.uvm.edu/process.php, verfügbar am 03.08.2011

[Chase, 2009]
Chase, Mark W.; Fay, Michael F.: *Barcoding of Plants and Fungi*. In: Science. - Washington: American Association for the Advancement of Science – 325(2009)5941, S. 862

[CBOL, 2011]
CBOL < cbolinfo@si.edu>: *What is DNA Barcoding?.* URL: < http://www.barcodeoflife.org/content/about/contact-cbol-secretariat>, verfügbar am 06-07-2011

[Cheng, 2008]
Cheng, Liang; Zhang, David Y.: *Molecular Genetic Pathology*. – 1. Auflage – New York: Humana Press, 2008, S. 106-115.

[Chevreux, 2005]
Chevreux, Bastien: *MIRA: An automated Genome and EST Assembler*. – 2005. – S. 17, 20f., Heidelberg, Ruprecht-Karls-University, Medizinische Fakultät, Doktorarbeit, 2005

[Dasenko, 2011]
Dasenko, Mark <mark@cgrb.oregonstate.edu>: Center for Genome Research & Biocomputing: *Troubleshooting*. URL:< http://corelabs.cgrb.oregonstate.edu/sequence/ troubleshoot>, verfügbar am 27.07.2011

[Ewing, 1998 a]
Ewing, Brent; Hillier, LaDeana; et. al.: *Base-Calling of Automated Sequencer Traces Using Phred: I. Accuracy Assessment*. In: Genome Research – New York: Cold Spring Harbor Laboratory Press – 8(1998)3, S. 175-185

[Ewing, 1998 b]
Ewing, Brent; Green, Phil: *Base-Calling of Automated Sequencer Traces Using Phred: II. Error Probabilities*. In: Genome Research – New York: Cold Spring Harbor Laboratory Press – 8(1998)3, S. 186-194

[Garcia-Hermosa, 2009]
Garcia-Hermosa, Dea; Hoinard, Damien; et. al: *Molecular and Phenotypic Evaluation of Lichtheimia corymbifera (Formerly Absidia corymbifera) Complex Isolates Associated with Human Mucormycosis: Rehabilitation of L. ramose*. In: Journal of clinical microbiology - Washington: American Society for Microbiology – 47(2009)12, S. 3862-3870

[Hansen, 2004]
Hansen, Andrea: Bioinformatik: *Ein Leitfaden für Naturwissenschaftler*. – 2. Auflage – Basel: Birkhäuser Verlag, 2004, S. 11f.

[Hajibabaej, 2005]
Hajibabaei, Mehrdad; deWaard, Jeremy R.; et. al.: *Critical factors for assembling a high volume of DNA Barcodes*. In: Phil. Trans. R. Soc. B – London: The Royal Society – B(2005)360, S. 1959 – 1967

[Hajibabaej, 2007]
Hajibabaei, Mehrdad; Singer, Gregory A.C.; et. al.: *DNA barcoding how it complements taxonomy, molecular phylogenetics and population genetics*. In: Trends in Genetics. - Cambridge: Cell Press- 23(2007)4, S. 167 - 172

[Hanner, 2009]
Hanner, Robert: < cbolinfo@si.edu>: *Data Standards for BARCODE Records in INSDC (BRIs)*. URL: < http://www.barcodeoflife.org/sites/default/files/legacy/pdf/DWG_data_ standards-Final.pdf>, verfügbar am 06-07-2011

[Hebert 2003]
Hebert, Paul D.N.; Cywinska, Alina; et. al.: *Biological identification through DNA Barcodes*. In: Proc. R. Soc. Lond. B – 270(2003)1512, S. 313-321

[Hebert, 2004]
Hebert, Paul D.N.; Penton, Erin H.; et. al.: *Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator*. In: Proc Natl Acad Sci U.S.A. - Washington : National Acedemy of Sciences - 101(2004)41, S. 14812-14817

[Hebert, 2005]
Hebert, Paul D. N.; Gregory, T. Ryan: The Promise of DNA Barcoding for Taxonomy. In: System Biology - Oxford: Society of Systematic Biologists – 54(2005)5, S. 852–859

[Heiner, 1998]
Heiner, Cheryl C.; Hunkapiller, Kathryn L.; et. al.: *Sequencing Multimegabase-Template DNA with BigDye Terminator Chemistry*. In: Genome Research. - New York: Cold Spring Harbor Laboratory Press – 8(1998)5, S. 557-561

[Ivanova, 2011]
Ivanova, Natalia V.; deWaard, Jeremy R.; et. al: nivanova@uoguelph.ca, jdewaard@uoguelph.ca: *Protocols for High-Volume DNA Barcode Analysis*. URL: http://www.barcodeoflife.org/sites/default/files/Protocols_for_High_Volume_DNA_Barcode_Analysis.pdf, verfügbar am 04.08.2011

[Janzen, 2009]
Janzen Daniel H.: *A DNA barcode for land plants*. In: Proc Natl Acad Sci U.S.A. - Washington: National Acedemy of Sciences – 106(2009)31, S. 12794-12797

[Kieleczawa, 2004]
Kieleczawa, Jan: *DNA Sequencing: Optimizing the Process and Analysis*. – 1. Auflage – London: Jones and Bartlett Publishers Publishers, 2004, S. 11 ff.

[Kress, 2008]
Kress, W. John; Erickson, David L.: *DNA barcodes: Genes, genomics, and bioinformatics*. In: PNAS. - Washington : National Acedemy of Sciences – 105(2008)8, S. 2761–2762

[Kress, 2009]
Kress, John W.; Erickson, David L.; et. al. <kressj@si.edu>: *Proposal to the consortium for the barcode of life for the adoption of a three-locus DNA-Barcode for land plants*. URL: <http://barcoding.si.edu/PDF/PlantWG/Kress%20PWG%20Proposal%20%209%20September%202009.pdf>, verfügbar am 04.08.2011

[Kropp, 2011]
Kropp, Doreen: Vergleichende Genanalyse zur Identifizierung humanpathogener, allergener und toxinogener Pilze. -2011. – Mittweida, Hochschule Mittweida, Mathematin/Naturwissenschaft/Informatik, Bachelor-Arbeit, 2011

[Lambert, 2005]
Lambert, D. M.; Baker, A.; et. al.: *Is a large-scale DNA-based inventory of ancient life possible*? In: Journal of Heredity. - Oxford: The American Genetic Association - 96(2005)3, S. 279–284

[Lawrence, 1993]
Lawrence, Charles B.; Solovyev, Victor V.: *Assignment of position-specific error probabiolity tp primary DNA sequence data*. In: Nucleic Acid Research. – Oxford: Oxford University Press – 22(1994)7, S.1272-1280

[Lipshutz, 1994]
Lipshutz, Robert J.; Taverner, Fred; et. al.: *DNA Sequence Confidence Estimation*. In: Genomics. - Massachusetts: Academic Press - 19(1994)3, S. 417-424

[Matz, 2005]
Matz, Mikhail V.;Nielsen, Rasmus.: *A likelihood ratio test for species membership based on DNA sequence data*. In: Phil.Trans. R. Soc. B - London: The Royal Society - 360(2005)1462, S. 1969–1974

[May, 2009]
May, Robert M.; Harvey, Paul H.: *Species Uncertainties*. In: Science - Washington: American Association for the Advancement of Science- 323(2009)5915, S. 687

[Merkl, 2009 a]
Merkl, Rainer; Waack, Stephan: *Bioinformatik Interaktiv: Grundlagen, Algorithmen, Anwendungen*. – 2. Auflage – Weinhein: WILEY-VHC, 2009, S. 261

[Merkl, 2009 b]
Merkl, Rainer; Waack, Stephan: *Bioinformatik Interaktiv: Grundlagen, Algorithmen, Anwendungen*. – 2. Auflage – Weinhein: WILEY-VHC, 2009, S. 482, 486f.

[Meyer, 2005]
Meyer, Christopher P.; Paulay, Gustav: *DNA Barcoding: Error Rates Based on Comprehensive Sampling*. In: PLoS Biology - Cambridge: Public Library of Science - 3(2005)12, S. e422

[Moritz, 2004]
Moritz, Craig; Cicero, Carla: *DNA Barcoding: Promises and Pitfalls*. In: PLoS Biology - Cambridge: Public Library of Science - 2(2004)10, S.1529-1531

[Nelson, 2005]
Nelson, David L.; Cox, Michael M.: *Lehninger Biochemie*. – 3. Auflage – Berlin Heidelberg: Springer Verlag, 2005, S. 370ff

[Nilsson, 2006]
Nilsson, R. Henrik; Ryberg, Martin; et. al: *Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective*. In: PLoS ONE. - Cambridge: Public Library of Science- 1(2006)1, S. 1-4

[Nölte, 2002]
Nölte, M.: *„Optimierung von Oligonukleotid-Bibliotheken für DNA-Mikroarrays“.* URL:< http://elib.suub.uni-bremen.de/publications/dissertations/E-Diss779_Dissertation_ Manfred _Noelte.pdf>, Universität Bremen, Mai 2002, verfügbar am 01.06.2011

[Nucleics, 2010]
Nucleics <information@nucleics.com>: *DNA Sequencing Troubleshooting*. URL:< http://www.nucleics.com/DNA_sequencing_support/DNA-sequencing-troubleshooting.html>, verfügbar am 27.07.2011

[Nüßlein, 2009]
Nüßlein, Bernhardv <info@nadicom.com>: *Nadicom informiert: Messe Newsletter der nadicom Gesellschaft für angewandte Mikrobiologie mbH*. URL:<http://www.nadicom .com/sites/default/files/files/Newsletter%20nadicom%201_2009.pdf>, verfügbar am 02.08.2011

[Primrose, 2006]
Primrose, S.B.; Twyman, Richard M.: *Principles of Gene Manipulation and Genomics*. – 7. Auflage – Oxford: Blackwell Publishing, 2006, S. 130

[Pruesse, 2007]
Pruesse, Elmar; Quast, Christian; et. al.: *SILVA: a comprehensive online-resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB*. In: Nucleic Acid Research. – Oxford: Oxford University Press – 35(2007)21, S. 7188-7196

[Rauch, 2007]
Rauch, P.: „Schimmelpilze in Wohngebäuden: Ursachen, Vermeidung und Sanierung“. – 1. Aufl. Leipzig: Verlag Peter Rauch, 2007, S. 16

[Ratnasingham, 2007]
Ratnasingham, Sujee van; Hebert, Paul D. N.*: BOLD: The Barcode of Life Data System (www.barcodeoflife.org)*. In: Molecular Ecology Notes. – Oxford: Blackwell Publishing – 7(2007)3, S. 355–364

[Richterich, 1998]
Richterich, Peter: *Estimation of Errors in “Raw” DNA Sequences: A Validation Study*. In: Genome Research. - New York: Cold Spring Harbor Laboratory Press – 8(1998)3 , S. 251- 259

[Ryberg, 2009]
Ryberg, Martin; Kristiansson, Erik; et. al.: *An outlook on the fungal internal transcribed spacer sequeces in GenBank and the introduction of a web-based tool for the exploration of fungal diversity*. In: New Phytologist. - Oxford: Wiley VHC – 181(2009)2, S. 471-477

[Savolainen, 2005]
Savolainen, Vincent; Cowan, Robyn S.; et. al*.: Towards writing the encyclopedia of life: an introduction to DNA-Barcoding*. In: Phil. Trans. R. Soc. B – London: The Royal Society – B(2005)360, 1805 – 1811

[Schmidt, 2001]
Krone, Andreas (Hrsg.): *Der Kammolch (Triturus cristatus): Verbreitung, Biologie, Ökologie und Schutz*. – 4. Sonderheft – Rangsdorf: Natur & Text, 2001, S. 179 – 191

[Sensen, 2002 a]
Sensen, Christoph W.: *Essentials of Genomics and Bioinformatics*. – 1. Auflage – Weinheim: Wiley-VHC, 2002, S. 167ff.

[Sensen, 2002 b]
Sensen, Christoph W.: *Essentials of Genomics and Bioinformatics*. – 1. Auflage – Weinheim: Wiley-VHC, 2002, S. 180

[Seifert, 2009]
Seifert, Keith A.: *Progress towards DNA barcoding of fungi*. In: Molecular Ecology Resource. – Oxford: Blackwell Publishing – 9(2009)1

[Steinke, 2006]
Steinke, Dirk; Brede, Nora: *Taxonomie des 21. Jahrhunderts: DNA-Barcoding*. In: Biologie in unserer Zeit. – Weinheim: Wiley-VCH – 36(2006)1, S. 40 – 46

[Stoeckle, 2003]
Stoeckle, Mark; Janzen, Daniel: *Taxonomy, DNA, and the Barcode of Life: Draft Conference Report*. In: Meeting held at Banbury Center – New York: Cold Spring Harbor Laboratory – September 10-12, 2003

[Stoeckle, 2004]
Stockle, Mark; Waggoner, Paul; Ausubel: *Barcoding of Life: ten reasons*. URL:< http://phe.rockefeller.edu/docs/TenReasonsBarcoding-1.pdf>,
verfügbar am 03.08.2011

[Stoeckle, 2008]
Stoeckle, Mark Y.; Hebert, Paul D.N.: *Barcode of Life*. In: Scientific American. – New York: Scientific American – 299(2008)4, S. 82 – 88

[Strandhagen, 2010]
Strandhagen, Ulrike: „Weiterentwicklung eines DNS-Microarrays zur Identifikation von humanpathogenen Pilzen“. URL:<http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-54311>, verfügbar am 18.05.2011

[Wendl, 2001]
Wendl, Michael C.; Korf, Ian; et. al.: *Automated Processing of Raw DNA Sequence Data*. In: IEEE Engineering in Medicine and Biology. – New York: IEEE - 20(2001)4, S. 41-48

[Wiemann, 1995]
Wiemann, Stephan; Stegemann, Josef; et. al: *Simultaneous On-Line DNA Sequencing on both Strands with two Fluorescent Dyes*. In: Analytical Biochemistry. - Massachusetts: Academic Press - 224(1995)1, S. 117-121

Anlagen

Anlagen, Datenbestand

Tabelle A-1: Zielorganismen mit Datenübersicht

Die Tabelle enthält alle Spezies-Namen der Pilze, die für die Entwicklung der Diagnostik beachtet werden sollen. Die Abkürzungen stehen stellvertretend für diese Namen. Des Weiteren wird angegeben, von welchen Spezies die ITS-Sequenzen durch die *Biotype Diagnostics GmbH* sequenziert wurden und ob eine Konsensus-Sequenz der Datenbanksequenzen vorliegt. Die letzten zwei Spalten geben wieder, welche Genom-Region in BOLD vorliegt und ob EPG vorhanden sind. Sind keine Daten zu dem entsprechenden Kriterium vorhanden, wurde das Feld leer gelassen.

Name	Abkürzung	Sequenziert	Konsensus	BOLD Sequenz	BOLD Trace
Absidia corymbifera	AbsCor	ja	ja		
Absidia glauca	AbsGla	ja	ja		
Acremonium alternatum	AcrAlt				
Acremonium kiliense	AcrKil	ja			
Acremonium murorum	AcrMur	ja		ITS	ja
Acremonium strictum	AcrStr				
Alternaria alternata	AltAlt	ja	ja		
Alternaria mali	AltMal		ja		
Alternaria citri	AltCit	ja	ja		
Alternaria tenuissima	AltTen	ja	ja		
Arthroderma benhamiae	ArtBen				
Aspergillus caesiellus	AspCae	ja			
Aspergillus candidus	AspCan		ja		
Aspergillus flavus	AspFla	ja	ja		
Aspergillus fumigatus	AspFum	ja	ja		
Aspergillus nidulans	AspNid		ja		
Aspergillus niger	AspNig		ja	COI	ja
Aspergillus ochraceus	AspOch		ja		
Aspergillus penicillioides	AspPen		ja		
Aspergillus restrictus	AspRes	ja			
Aspergillus sydowii	AspSyd		ja		
Aspergillus tamarii	AspTam		ja		
Aspergillus terreus	AspTer		ja		
Aspergillus ustus	AspUst		ja		
Aspergillus versicolor	AspVer		ja		
Aureobasidium pullulans var melanigeum	AurPulVMel		ja		
Aureobasidium pullulans var pullulans	AusPul	ja	ja		
Botrytis cinerea	BotCin	ja	ja		

Candida (Pichia) guilliermondii	CanGui		ja		
Candida albicans	CanAlb		ja		
Candida glabrata	CanGla		ja	COI	nein
Candida krusei (Issatchenkia orientalis)	CanKru		ja		
Candida parapsilosis	CanPar		ja	COI	nein
Candida tropicalis	CanTro		ja		
Chaetomium brasiliense	ChaBra	ja			
Chaetomium globosum	ChaGlo	ja	ja		
Chaetomium murorum	ChaMur	ja			
Cladosporium carrionii (syn. Cladophialophora)	ClaCar		ja		
Cladosporium cladosporioides	ClaCla	ja	ja		
Cladosporium herbarum	ClaHer	ja	ja		
Cladosporium sphaerospermum	ClaSph	ja	ja		
Engyodontium album	EngAlb	ja	ja		
Epidermophyton floccosum	EpiFlo		ja	COI	nein
Eurotium amstelodami	EurAms	ja	ja		
Eurotium chevalieri	EurChe	ja			
Eurotium rubrum	EurRub	ja	ja		
Exophiala dermatitidis	ExoDer	ja			
Fusarium culmorum	FusCul	ja			
Fusarium oxysporum	FusOxy	ja	ja	ITS, COI	ja, nein
Fusarium solani	FusSol	ja	ja	ITS, COI	ja, ja
Fusarium verticillioides	FusVer	ja			
Geotrichum candidum	GeoCan	ja		ITS	ja
Geomyces pannorum	GeoPan	ja		ITS	ja
Hormoconis resinae	HorRes	ja			
Malassezia furfur	MalFur		ja		
Malassezia globosa	MalGlo		ja		
Microsporum canis	MicCan		ja		
Microsporum gypseum	MicGyp		ja		
Mucor plumbeus	MucPlu	ja			
Mucor racemosus	MucRac	ja	ja		
Oidiodendron griseum	OidGri	ja			
Paecilomyces lilacinus	PaeLil			ITS	ja
Paecilomyces variotii	PaeVar				
Penicillium brevicompactum	PenBre	ja		ITS, COI	nein, ja
Penicillium chrysogenum (syn. notatum)	PenChr	ja	ja	COI	ja
Penicillium aurantiogriseum (aurantiovirens)	OenAur			COI	ja

Penicillium brevicompactum	PenBre	ja		ITS, COI	nein, ja
Penicillium chrysogenum (syn. notatum)	PenChr	ja	ja	COI	ja
Penicillium commune	PenCom		ja	COI	ja
Penicillium digitatum	PenDig			COI	ja
Penicillium expansum	PenExp		ja	COI	ja
Penicillium funiculosum	PenFun		ja		
Penicillium glabrum	PenGla		ja	ITS, COI	ja, ja
Penicillium griseofulvum	PenGri	ja	ja	COI	ja
Penicillium marneffei	PenMar	ja		COI	nein
Penicillium piceum	PenPic	ja			
Penicillium purpurogenum	PenPur				
Penicillium verruculosum	PenVer	ja			
Phialophora fastigiata	PhiFas	ja			
Phialophora heteromorpha	PhiHet				
Phoma glomerata	PhoGlo				
Rhizopus oryzae	RhiOry	ja		COI	nein
Rhizopus stolonifer	RhiSto	ja			
Rhodotorula glutinis	RhoGlu		ja		
Rhodotorula mucilaginosa	RhoMuc		ja		
Saccharomyces cerevisiae	SacCer		ja	COI	nein
Scopulariopsis brevicaulis	ScoBre				
Scopulariopsis fusca	ScoFus	ja			
Stachybotrys chartarum	StaCha				
Syncephalastrum racemosum	SynRac	ja			
Trichoderma harzianum (Hypocrea lixii)	TriHar	ja		ITS	ja
Trichoderma viride	TriVir	ja		18S, 28S	nein, nein
Trichophyton erinacei	TriEri				
Trichophyton interdigitale (anthropophil)	TriIntA				
Trichophyton interdigitale (zoophil)	TriIntZ				
Trichophyton rubrum	TriRub		ja	COI	nein
Trichophyton tonsurans	TriTon		ja		
Trichophyton verrucosum	TriVer		ja		
Trichophyton violaceum (soudanense)	TriVio		ja		
Trichosporon cutaneum	TriCut		ja		
Verticillium /Lecanicillium lecanii	VerLec				
Wallemia sebi	WalSeb	ja	ja		

Tabelle A-2: Projekte mit Pilz-Kontext in BOLD und deren Inhalt

Die Tabelle gibt die in BOLD enthaltenen Daten pro Projekt wieder. Die erste und zweite Spalte geben jeweils das Kürzel für ein Projekt und die Gen-Region, über die Daten enthalten sind, an. Die nächsten drei Spalten geben eine Übersicht der Gesamtzahl der vorhandenen Traces, Sequenzen und Spezies. Zuletzt wird die Anzahl der nutzbaren Daten für das Projekt der *Biotype Diagnostics GmbH* abenfalls aufgeschlüsselt in Traces, Sequenzen und Spezies wiedergegeben. Die in den Klammern angegebenen Symbole ordnen die Daten den jeweiligen Genomregionen (wenn im Projekt mehr als eine enthalten ist) zu. Dabei steht I für ITS, C für COI, 18 für 18S und 28 für 28S

Allgemeines		Gesamtanzahl der in BOLD vorhandener			In Bezug auf die Zielspezies enthaltenen		
Projekt Code	Genom-Region	Traces	Sequenzen	Spezies	Traces	Sequenquen-zen	Spezies
NTDMF	18S, 28S		1(18), 1(28)	1(18), 1(28)		1(18) 1(28)	1
CND	ITS	975	798	344			
CEFI	ITS	274	69	14	48	11	3
CAM	ITS	96	44	32			
CHBAR	ITS, COI, 28S		107(I), 23(C), 83(28)	16(I), 9(C), 16(28)			
COBAR	ITS		19	4			
CRBAR	ITS		82	9			
MUSH	COI	56	65	55			
YYY	ITS, COI	110(I), 41(C)	167(I), 167(C)	103(I), 103(C)			
FRBAR	ITS		14	6			
GBF	ITS	508	223	175	15	4	3
GMBAR	ITS		8	6			
HY	COI	117	67	30	6	2	1
HTWO	COI	14	7	7			
LHM	ITS, COI	114(I), 121(C)	52(I), 52(C)	9(I), 9(C)			
MEBAR	ITS, COI, 28S	195(I)	347(I), 33(C), 333(28)	30(I), 8(C), 28(28)			
PATE	ITS	115	54	31	18	9	2
PSP	COI	762	354	70	86	40	9
PUBAR	ITS, COI		23(I), 23(C)	14(I), 13(C)			
SAM	ITS	271	199	118			
SEA	ITS	250	190	92			
USA	ITS	380	224	142			
RUBAR	ITS		69	23			
WSF	ITS	723	353	47	70	35	4
GBFB	COI		56	56		10	9

Tabelle A-3: Klassifizierung der Sequenzen und Auflistung benötigter Parameter

Die Tabelle stellt die einzelnen Klassifizierungen zu jeder Sequenz in Kombination der für die Einteilung benötigten Parameter dar. Es wird die Länge (Anzahl Basen), der Anteil qualitativ sehr guter Basen (PHRED > 30) sowie schlechter Basen (PHRED < 20) und der Mittelwert aller PHRED-Werte der Sequenz aufgelistet. Anhand dieser Werte wurde die Klassifizierung (vgl. Kapitel 6.1.3) vorgenommen. Eine Sequenz wird in eine der vier Kategorien „failed“, „low“, medium“ und „high“ eingeordnet. Die Ergebnisse sind in Kapitel 6.3 zusammengefasst dargestellt.

Name	Anzahl Basen	PHRED >30 (%)	PHRED < 20 (%)	PHRED Mittelwert	Klassifizierung
AbsCor F	792	96,338	2,273	60,1	medium
AbsCor R	790	98,481	0,886	59,6	medium
AbsGla F	593	60,034	19,562	31,9	low
AbsGla R	581	91,394	4,991	57,8	medium
AcrKil F	522	95,594	1,724	62,2	medium
AcrKil R	530	96,038	2,264	62,1	medium
AcrMur F1	500	97,600	1,800	59,7	medium
AcrMur F2	515	87,573	4,660	50,0	medium
AcrMur R1	522	98,084	1,533	65,0	medium
AcrMur R2	531	98,682	0,753	62,8	medium
AltAlt F	510	98,235	0,392	61,8	medium
AltAlt R	496	98,185	1,210	61,7	medium
AltCiT F	488	94,672	3,893	60,9	medium
AltCit R	505	97,624	0,792	64,4	medium
AltTen F	515	97,476	0,583	61,8	medium
AltTen R	515	95,922	1,553	63,8	medium
AspCae F	550	94,909	1,636	60,3	medium
AspCae R	528	97,159	1,326	64,0	medium
AspFla F	548	95,620	1,825	61,8	medium
AspFla R	523	97,514	1,338	64,1	medium
AspFum F	269	67,658	16,729	34,7	low
AspFum R	338	83,728	8,284	42,0	medium
AspRes F	550	95,455	2,364	61,3	medium
AspRes R	535	93,084	5,047	60,5	medium
AurPul F	522	96,360	1,533	63,4	medium
AurPul R	504	98,214	1,190	65,1	medium
BotCin F	479	97,495	1,253	58,3	medium
BotCin R	463	96,544	1,944	57,3	medium
ChaBra F	515	95,922	1,165	62,3	medium
ChaBra R	508	97,638	0,787	64,1	medium
ChaGlo F	522	97,126	0,383	61,2	medium
ChaGlo R	515	97,087	1,553	63,4	medium
ChaMur F1	412	36,165	33,010	26,4	low

ChaMur F2	1	0	100	6,0	failed
ChaMur R1	1	0	100	10,0	failed
ChaMur R2	1	0	100	18,0	failed
ClaCla F	484	80,372	12,810	46,1	medium
ClaClaR	519	94,412	2,119	56,8	medium
ClaHer F	491	96,334	1,426	61,7	medium
ClaHer R	503	95,626	1,789	63,0	medium
ClaSph F	488	97,746	1,230	64,7	medium
ClaSph R	492	96,545	1,423	62,6	medium
EngAlb F	499	94,389	2,004	49,6	medium
EngAlb R	504	79,167	7,143	38,4	low
EurAms F	502	97,211	1,394	63,2	medium
EurAms R	474	98,101	1,266	64,5	medium
EurChe F	501	95,210	1,796	62,1	medium
EurChe R	474	98,312	1,266	64,6	medium
EurRub F	509	95,678	0,589	60,7	medium
EurRub R	491	96,334	2,240	63,3	medium
ExoDer F	591	97,293	1,015	64,0	medium
ExoDer R	582	98,454	1,203	64,4	medium
FusCul F	1	0	100	8,0	failed
FusCul R	499	91,583	6,012	51,7	medium
FusOxy F	465	98,280	1,290	63,3	medium
FusOxy R	473	97,463	1,480	64,2	medium
FusSol F	506	97,826	0,988	63,8	medium
FusSol R	447	61,074	21,477	35,4	low
FusVer F	480	97,292	1,250	63,3	medium
FusVer R	488	96,311	1,025	63,1	medium
GeoCan F	309	97,087	1,618	60,3	medium
GeoCan R	306	96,078	2,288	62,3	medium
GeoPan F	520	96,154	2,115	58,4	medium
GeoPan R	500	92,400	4,400	59,1	medium
HorRes F	489	97,342	2,045	63,5	medium
HorRes R	482	96,680	1,660	63,8	medium
MucPlu F	585	97,607	1,197	61,8	medium
MucPlu R	561	95,900	1,783	63,4	medium
MucRac F	594	95,623	1,852	60,1	medium
MucRac R	552	97,464	1,449	57,5	medium
OidGri F	1	0	100	14,0	failed
OidGri R	1	0	100	8,0	failed
PenBre F	464	88,793	7,328	50,1	medium
PenBre R	521	96,353	2,495	49,9	medium
PenChr F	535	96,449	1,682	63,2	medium
PenChr R	526	97,529	1,901	63,7	medium

PenGri F	512	89,258	6,836	48,9	medium
PenGri R	530	94,151	3,019	52,2	medium
PenMar F	518	97,297	0,579	59,5	medium
PenMar R	515	97,087	1,553	64,0	medium
PenPic F1	1	0	100	10,0	failed
PenPic F2	1	0	100	14,0	failed
PenPic R1	1	0	100	6,0	failed
PenPic R2	1	0	100	8,0	failed
PenVer F	470	99,149	0,426	63,0	medium
PenVer R	521	95,777	1,919	61,9	medium
PhiFas F	526	98,669	0,760	61,8	medium
PhiFas R	525	97,333	1,524	64,0	medium
RhiOry F	585	95,043	2,393	56,9	medium
RhiOry R	566	98,233	1,237	64,1	medium
RhiSto F	501	23,353	56,088	23,3	low
RhiSto R	308	64,935	22,403	43,7	medium
ScoFus F	1	0	100	7,0	failed
ScoFus R	1	0	100	8,0	failed
SynRac F	527	32,448	34,535	27,8	low
SynRac R	450	90,222	6	57,6	medium
TriHar F	469	11,727	57,143	19,2	failed
TriHar R	6	0	50	21,7	failed
TriVir F	522	98,084	1,149	64,0	medium
TriVir R	539	97,032	1,484	64,7	medium
WalSeb F	526	80,608	7,605	47,4	medium
WalSeb R	500	83,600	5,800	44,8	medium

Tabelle A-4: Irregularitäten der Traces

In der Tabelle sind alle Irregularitäten, welche in den Elektropherogrammen beobachtet wurden, aufgelistet. Die Bezeichnungen setzen sich aus den Abkürzungen der jeweiligen Spezies und ein F für Forward oder ein R für Reverse zusammen. Ziffern nach einem R oder F geben bei *AcrMur, ChaMur* und *PenPic* an, um welchen der beiden Forward- oder Reverse-Traces es sich handelt (1 wenn Dateiendung „ITS1.ITS1_**1**.ab1“ oder „ITS4.ITS4_**1**.ab1“ / 2 wenn „ITS2W.ITS1-**2W**.ab1“ oder „ITS4.ITS4_**2**.ab1“).

Irregularität	Trifft zu auf	Bemerkung
Slippage	RhiSto F+R	
Dye Blob zwischen 80-120 nt	ChaMur F2, ClaCla F, FusCul F+R, OidGri F, PenGri F+R, PenPic F2, RhiSto F, ScoFus F, TriHar F	Dye Blob der Base C
	AltCit F, GeoPan R	Dye Blob der Base T
	AspFum F+R, AspRes R, ChaMur F1+R2+R1, OidGri R, PenBre F PenPic F1+R2+R1, ScoFus R, TriHar R	Mehrere überlagerte Dye Blobs
	AbsGla F, AspCae F, BotCin R, ChaGlo R, ClaHer F + R, ClaCla R, ClaSph R, EngAlb R, FusSol R, GeoPan F, PenBre R, RhiOry F, WalSeb F	Nur schwache Dye Blobs
Geringe Signalstärke	AspFum F, ChaMur F1+F2+R1+R2, FusCul F, OidGri F+R, PenGri R, PenPic F1+F2+R1+R2, ScoFus F+R, TriHar F+R	Stark signalschwach
	AspFum R, ClaCla F, FusCul R, FusSol R (ab Base 150 wird Signal plötzlich schwächer) PenBre F, PenGri F, RhiSto F	signalschwach
Chimären	SynRac R (ab Base 430) SynRac F (ab Base 130)	Zu Anfang klar erkennbare Peaks
Fehlerhafte Primer-synthese	AbsGla F: N-1 WalSeb R: N-1	N-1 bedeutet Peaks verschoben nach links
Konta-mination	Eng Alb F+R, AcrMur F2: N+1 Shift (nur am Anfang)	Bei EngAlb F nur ganz schwach

Tabelle A-5: Assemblierung vor manueller Annotation

Aus allen Assemblierungen wurden die Anzahl der Lücken, Mismatches und Ns bestimmt. Als Lücken werden Insertionen und Deletionen bezeichnet. Einige Sequenzen konnten keine Konsensus hervorbringen, andere wiederum waren auch ohne manuelle Korrektur fehlerfrei.

Spezies	Lücken	Mismatches	Ns	Spezies	Lücken	Mismatches	Ns
AbsGla	1	6		EurChe	1		
AcrKil	3	0		EurRub	2	1	
AcrMur	2	1		ExoDer	1		
AltCit		3		FusSol		5	3
AltTen	2			FusVer	1	1	
AspCae	2		1	GeoCan	1		
AspFla	1			GeoPan	3	4	
AspFum		2		HorRes		1	
AspRes	1	3		PenBre		10	
AurPul	2			PenChr	2		
ChaBra	1			PenGri		11	5
ChaGlo	1			PhiFas	1		
ClaCla	1	12	16	RhiOry	1	1	
ClaHer	2			SynRac	1	11	
ClaSph	1	1		TriVir	1		
EngAlb	1	4	4	WalSeb	2	1	
EurAms		1					
Keine Konsenus: ChaMur, OidGri, PenPic, ScoFus, TriHar, FusCul, RhiSto							
Fehlerfreie Konsensus: AbsCor, AltAlt, BotCin, FusOxy, MucPlu, MucRac, PenMar, PenVer							

Tabelle A-6: Assemblierung nach der manuellen Annotation

Die Tabelle gibt eine Übersicht der Lücken, Mismatches und Ns in den Konsensen nach der manuellen Korekktur. Es sind nur die Spezies dargestellt, bei denen noch Lücken, Mismatches oder Ns in den Konsensen auftreten, die auch manuell nicht bereinigt werden konnten.

Spezies	Nach manuellen Annotation		
	Lücken	Mismatches	Ns
AbsGla		4	
AcrMur		1	
EngAlb		2	
FusVer		1	
PenGri		2	
RhiOry		1	
SynRac	1	10	

Tabelle A-7: Übersicht der für die Validierung der Klassifizierung genutzten BOLD-Sequenzen

Die Tabelle gibt alle aus BOLD benutzten Sequenzen für die Überprüfung der Qualitätskriterien wieder. Es sind die sieben Spezies, von denen sowohl Daten als BOLD als auch sequenzierte Daten vorliegen. Angegeben ist das Projekt aus den sie stammen mit zugehöriger ID und Länge. Das Feld „Barcode“ gibt an, ob die Sequenz potentiell als Barcode geeignet ist. Ebenfalls ist die Anzahl der Traces in den jeweiligen Klassifizierungen und deren Gesamtzahl angegeben (H = high, M = medium, L = low, F = failed, NT = kein Trace zu der Sequenz vorhanden).

Spezies	Projekt	Process ID	Sequenz Länge	Barcode	Anzahl Traces mit					Anzahl Traces gesamt
					H	M	L	F	NT	
AcrMur	GBF	GBF183-08	531[0n]	nein	3	1				4
FusOxy	CEFI	CEFI008-09	556[0n]	Nein	3		1			4
	CEFI	CEFI014-09	549[0n]	Nein	3	1				4
	CEFI	CEFI016-09	552[0n]	Nein	3		1			4
	CEFI	CEFI022-09	545[0n]	Nein	4					4
	CEFI	CEFI023-09	526[0n]	Nein	3	3	2			8
	CEFI	CEFI032-09	516[0n]	Nein	3		1			4
	CEFI	CEFI037-09	528[0n]	Nein	1	2		1		4
	WSF	WSF119-09	548[0n]	Nein	2					2
	WSF	WSF024-09	557[0n]	Nein	2					2
	WSF	WSF035-09	548[0n]	Nein	2					2
	WSF	WSF036-09	547[0n]	Nein	2					2
	WSF	WSF042-09	588[0n]	Nein	2					2
	WSF	WSF059-09	531[0n]	Nein				2		2
	WSF	WSF065-09	485[0n]	Nein	1		1			2
	WSF	WSF072-09	560[0n]	Nein	2					2
	WSF	WSF074-09	547[1n]	Nein	2					2
	WSF	WSF102-09	548[0n]	Nein	2					2
	WSF	WSF094-09	556[0n]	Nein		2				2
	WSF	WSF104-09	584[0n]	Nein	2					2
	WSF	WSF129-09	547[0n]	Nein	2					2
FusSol	CEFI	CEFI025-09	546[0n]	Nein	2	2				4
	WSF	WSF110-09	550[0n]	Nein		1	1			2
	WSF	WSF044-09	570[0n]	Nein	2					2
	WSF	WSF066-09	573[2n]	Nein		2				2
	WSF	WSF089-09	569[0n]	Nein	2					2
GeoPan	GBF	GBF130-08	514[0n]	Nein	1	1				2
GeoCan	GBF	GBF172-08	565[0n]	Nein	1	1		3		5
PenBre	PATE	PATE005-08	605[0n]	Nein					1	0
TriHar	CEFI	CEFI002-09	621[0n]	Nein	4					4
	CEFI	CEFI036-09	615[0n]	Nein	1	1	2			4
	WSF	WSF069-09	617[0n]	Nein	2					2
	WSF	WSF095-09	628[0n]	Nein	2					2

Tabelle A-8: Ergebnis des Vergleichs der assemblierten Datenbank-Sequenzen mit Referenz-Sequenzen

Es sind die 23 Spezies für die die vorliegenden assemblierten Datenbanksequenzen hinsichtlich ihrer Genauigkeit untersucht worden. Die verwendeten referenz-Sequenzen stammen entweder aus der BOLD oder von den Sequenzierungen. Lücken und Mismatches wurden in zwei Bereichen (die vermutlichen ITS –Sequenzen) getrennt gezählt. Die Grenzen dieser Bereiche wurden nach der größten Teil-Sequenz ohne Mismatches oder Lücken ausgesucht (Bereich angegeben unter „Bereich ohne Fehler“). Der Assemblierungs-Bereich gibt an, an welcher Stelle die Referenz-Sequenz mit der assemblierten Datenbanksequenz übereinstimmt.

Spezies	Herkunft Referenz	Lücken (ITS1)	Mismatches (ITS1)	Lücken (ITS2)	Mismatches (ITS2)	Fehler insgesamt	Assemblierungs-Bereich	Bereich ohne Fehler
AbsCor	Sequenzierung	1	6	5	8	20	3174-4010	3467-3682
AbsGla	Sequenzierung	5	35	15	26	81	716-1356	957-1111
AltAlt	Sequenzierung	4	8	6	8	26	1339-1896	1510-1700
AltCit	Sequenzierung	0	0	0	0	0	348-919	0
AltTen	Sequenzierung	2	0	0	1	3	859-1420	922-1413
AspFla	Sequenzierung	2	0	0	1	3	620-1207	688-1207
AspFum	Sequenzierung	0	5	0	3	8	1824-2294	1879-2256
AurPul	Sequenzierung	0	8	0	1	9	591-1151	766-1151
BotCin	Sequenzierung	0	0	0	1	1	2320-2851	2320-2849
ChaGlo	Sequenzierung	0	0	0	3	3	3178-3750	1813-3750
ClaCla	Sequenzierung	0	10	0	0	10	1797-2364	1813-3750
ClaHer	Sequenzierung	1	9	0	0	10	1661-2212	1721-2212
ClaSph	Sequenzierung	0	0	0	1	1	1726-2260	0
EngAlb	Sequenzierung	28	24	19	41	112	1713-2313	1952-2050
EurAms	Sequenzierung	1	0	0	2	3	867-1413	875-1404
EurRub	Sequenzierung	0	0	0	2	2	1391-1945	1391-1845
FusOxy	BOLD, Sequenzierung	1	0	1	1	3	1724-2296	1881-2134
FusSol	BOLD, Sequenzierung	0	3	2	6	11	3233-3823	3427-3645
MucRac	Sequenzierung	3	29	0	18	50	3487-4116	3702-3901
PenChr	Sequenzierung	0	0	1	1	2	587-1165	587-1013
PenGla	BOLD	0	1	0	0	1	1745-2346	1787-2346
PenGri	Sequenzierung	0	3	0	0	3	1768-2357	1770-2357
WalSeb	Sequenzierung	3	6	6	6	21	1410-1974	1498-1607

Anlagen, Quellcode

```
import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.File;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.util.ArrayList;


public class ReadQualities
{
	private ArrayList<String> base = new ArrayList<String>();
	private ArrayList<String> phred = new ArrayList<String>();
	private String [] fileNames;

	//Einlesen des angegebenen Files

	public void readFile(String fileName) throws IOException
	{
		base.clear();
		phred.clear();
		String line = "";
		boolean begin_DNA = false;
		boolean end_DNA = false;
		File file = new File(fileName);
		BufferedReader buf = new
			BufferedReader(new FileReader(file));
		while((line = buf.readLine()) != null )
		{
			if(line.contains("BEGIN_DNA")) begin_DNA = true
			if(line.contains("END_DNA")) end_DNA = true;
			if(begin_DNA == true && end_DNA ==
				false && !line.contains("BEGIN_DNA"))
			{
				base.add(line.split(" ")[0]);
				phred.add(line.split(" ")[1]);
			}
		}
		buf.close();
	}
		//Lesen aller Dateien in einem Ordner
	public void readDirectoryFiles(File directory)
	{  fileNames = directory.list();}

	//Schreiben des Inhaltes aller Dateien im Ordner in eine
	Textdatei
	public void writeFile(String fileName, String sequence)
		throws IOException
	{
		BufferedWriter buf = new BufferedWriter(
			new FileWriter(new File(fileName)));
		buf.write(sequence);
		buf.close();
	}
```

```
	public ArrayList<String> getBase()
	{ return base;}

	public ArrayList<String> getPhred()
	{ return phred; }

	public String [] getFileNames()
	{return fileNames;}

	public static void main(String[] args)
	{
		ReadQualities qual = new ReadQualities();
		qual.readDirectoryFiles(
			new File("D:\\Anwendungen\\Kommunikation\\" +
			"Bachelor\\CodonCodeAligner\\Projects\\ Pilze nachse
			quenziert\\phd_dir"));
		String [] fileNames = qual.getFileNames();
		String fileContent = "";
		for(int i = 0; i < fileNames.length; i++)
		{
			try
			{
				qual.readFile("D:\\Anwendungen\\Kommunikation\\"
				 +"Bachelor\\CodonCodeAligner\\Projects\\Pilze
				 nachsequenziert\\phd_dir\\" + fileNames[i]);
				fileContent = fileContent + "\n\n" +
				 fileNames[i] + " , " + qual.getBase() + "\n"  +
				 ", " + qual.getPhred();

			}
			catch(IOException e)
			{e.printStackTrace();}

		}
		fileContent=fileContent.replace('[', ' ').replace(']',' ');
		try
		{
			qual.writeFile("C:\\Users\\Bianca\\Desktop\\Bachelor
			\\CodonCodeAlignerAnalyse\\a.txt", fileContent);
		}
		catch (IOException e)
		{
			e.printStackTrace();
		}
	}
}
```

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.
Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.
Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, den 20.August 2011

_____

Bianca Liebscher